**Ana Filipa de Almeida Januário**

# Modelos estatísticos para prognóstico em recém nascidos prematuros extremos


**Departamento de Matemática**
**Faculdade de Ciências da Universidade do Porto**
**setembro de 2012**

**Ana Filipa de Almeida Januário**

# Modelos estatísticos para prognóstico em recém nascidos prematuros extremos


*Tese submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Engenharia Matemática*

Orientação Científica FCUP
Orientadora: Prof.ª Doutora Sónia Gouveia
Coorientador: Prof. Doutor Joaquim Pinto da Costa

Orientação MJD-CHP
Coorientadora: Dr.ª Isabel Sá

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
setembro de 2012

# Agradecimentos

Apesar do caráter individual inerente à realização de uma Tese de Mestrado, esta só se tornou possível com o apoio e colaboração de diversas pessoas e instituições. A todas elas, não posso deixar de expressar o meu sincero agradecimento.

À Prof.ª Doutora Sónia Gouveia pela excelente orientação, apoio e disponibilidade demonstrados ao longo da execução deste trabalho. Agradeço ainda todas as sugestões, conselhos e ensinamentos decorrentes da sua experiência e que em muito valorizaram esta investigação. Obrigada pelo estímulo constante, pela consideração e demonstração de confiança no nosso trabalho.

Ao Prof. Doutor Joaquim Pinto da Costa pela disponibilidade e partilha de conhecimentos.

Ao Centro Hospitalar do Porto por me ter acolhido e pela oportunidade que me concedeu de encetar a realização deste estudo.

À equipa médica da Maternidade de Júlio Dinis: Dr.ª Isabel Sá, Dr. Miguel Fonte, Dr.ª Alexandra Almeida, Dr.ª Carmen Carvalho, Dr. Joaquim Saraiva e Dr.ª Paula Soares pela cedência dos dados e pelas preciosas discussões médicas, fruto dos seus conhecimentos e experiência e que em muito enriqueceram o meu trabalho. Um agradecimento especial à Dr.ª Isabel Sá e à Dr.ª Alexandra Almeida pelo enorme interesse, disponibilidade e por me permitirem conhecer melhor o mundo da neonatolgia.

Ao Gabinete de Estatística, Modelação e Aplicações Computacionais e ao Centro de Matemática da Universidade do Porto agradeço o profissionalismo e o excelente ambiente de trabalho proporcionado. Este estudo foi parcialmente financiado pelo GEMAC/CMUP, financiados pelo FEDER através do programa COMPETE (ref. FCOMP-01-0124-FEDER-022656, CMUP, www.fc.up.pt/cmup/) e pelo laboratório Abbot.

À minha família e a todos os meus amigos, por todo o apoio, motivação, companheirismo e momentos de distração durante esta etapa.

À minha avó Mimi, simplesmente por ser o meu ídolo e fonte de inspiração desta caminhada.

Aos meus pais gostaria de expressar a enorme gratidão, por todo o carinho e apoio incondicional demonstrados em todos os momentos. Agradeço ainda a oportunidade que me concederam de enriquecer a minha formação académica, o incentivo indescritível e a compreensão nos momentos de ausência durante esta jornada.

À minha irmã Filipa, pela cumplicidade e por ter sido ouvinte atenta de cada página desta Tese.

Ao André agradeço a presença em todos os momentos, as palavras de incentivo e encorajamento e o ânimo com que me presenteou ao longo desta caminhada.

A todos quantos não citados individualmente, mas que se sabem credores de desmedido reconhecimento, o meu sincero agradecimento.

# Resumo

Nos cuidados a recém nascidos prematuros extremos, o diagnóstico/prognóstico célere e preciso, seguido de intervenção terapêutica rápida pode significar a diferença entre a vida e a morte ou, em caso de sobrevivência, minorar efeitos colaterais, aumentando assim as hipóteses de uma vida sem sequelas major. Existem, portanto, duas preocupações cruciais: reduzir a probabilidade de mortalidade e de morbilidade destes recém nascidos delicados. No entanto, avanços na terapêutica e na intervenção médica obrigam a um olhar retrospetivo que permita recolher dados e também um olhar prospetivo que permita avaliar o desenvolvimento neurológico, o que implica estudos prolongados no tempo. É neste contexto que se enquadram os contributos desta dissertação.
Neste trabalho são analisados dados de 205 recém nascidos prematuros extremos (<27 semanas de gestação), seguidos na Unidade Maternidade de Júlio Dinis - Centro Hospitalar do Porto (MJD-CHP) entre os anos 2000 e 2009. Cada recém nascido foi seguido num período máximo de dois anos: ao final do primeiro ano avaliou-se a mortalidade (sim/não) e, no caso de sobrevivência, ao final do segundo ano de vida avaliou-se a existência de sequelas major (morbilidade, sim/não). Esta investigação tem dois objetivos essenciais: por um lado, identificar fatores de risco precoces e, por outro, construir modelos preditivos de mortalidade e de morbilidade. Para o efeito, foram comparados modelos baseados em regressão logística e árvores de decisão, metodologias amplamente utilizadas na literatura da especialidade e que permitem obter informação complementar na análise dos dados. Especificamente, a regressão logística permite estimar a probabilidade de "sucesso"e as árvores de decisão permitem definir regras unívocas de decisão, representando-se de uma forma muito intuitiva e adequada à prática clínica.
Os resultados deste trabalho evidenciam a idade gestacional baixa como fator de risco muito determinante de mortalidade. Adicionalmente, foram também identificados como significativos, o baixo peso e o nascimento em hospital menos diferenciado com necessidade de transferência para tratamento. A precisão global de classificações corretas dos modelos de mortalidade foi, aproximadamente, de 70%. As conclusões para o estudo de morbilidade foram limitadas pelo tamanho reduzido da amostra, embora a gestação múltipla tenha sido identificada como um potencial fator de risco. As conclusões obtidas para os recém nascidos seguidos na MJD-CHP, justificam continuar a recolha de dados assim como implementar um estudo mais alargado para caraterizar a realidade portuguesa.
O famoso estatístico George Box (1987) sintetiza todo o estudo ao citar que *"Essentially, all models are wrong, but some are useful"*. De facto, embora estes modelos preditivos sejam aproximações da realidade, espera-se que os resultados decorrentes desta investigação tenham elevada repercussão e impacto na prática clínica. Em particular, modelos preditivos como forma de auxiliar de prognóstico clínico são contributos indispensáveis para uma resposta mais adequada às preocupações dos pais e para uma melhor alocação de recursos, quer humanos quer técnicos, nas unidades de cuidados neonatais.

**Palavras-chave:** MODELOS PREDITIVOS, REGRESSÃO LOGÍSTICA, ÁRVORES DE DECISÃO, RECÉM NASCIDOS PREMATUROS EXTREMOS, MORTALIDADE, MORBILIDADE.

# Abstract

In the care of extremely premature newborns, a fast and accurate diagnosis/prognosis, followed by an immediate therapeutic intervention can make the difference between life and death or, in case of survival, minimize side effects and increase the chances of a life without major sequels. Therefore, there are two crucial concerns: reduce the probability of mortality and of morbidity of these delicate newborns. However, advances in medical therapy and intervention demand a retrospective view, that allowing to collect data, and also a prospective view to allow the assessment of the neurological development, which involves time extended studies. It is in this context that this dissertation makes its contributions.

This work considers data from 205 extremely premature newborns (<27 weeks of gestation), followed at the Unit Maternidade de Júlio Dinis - Centro Hospitalar do Porto (MJD-CHP) between 2000 and 2009. Each newborn was followed during a maximum period of two years: Mortality (yes/no) was evaluated at the end of the first year while, in case of survival, the existence of major sequels (morbidity, yes/no) was evaluated at the end of the second year of life. This investigation has two main objectives: firstly, to identify early risk factors and, secondly to build predictive models of mortality and of morbidity. To this end, models based on logistic regression and decision trees were compared, both methodologies are widely used in the literature and allow to obtain complementary information concerning the data analysis. Specifically, logistic regression allows to estimate the "success" probability and decision trees allows to define univocal decision rules, with possible intuitive representation, very suitable for daily clinical practice.

The results of this work show that low gestational age is a very important risk factor that determines mortality. Additionally, low weight and the birth in less specialized hospitals (with the need of transfer for treatments) were identified as significant. The overall accuracy of correct mortality classification models was approximately 70%. Regarding the morbidity study, the conclusions were limited by the reduced sample size , although multiple pregnancies have been identified as a potential risk factor. The conclusions obtained for the newborns followed on the MJD-CHP, justify to further continue data collection at MJD as well as to implement a larger study to characterize the Portuguese reality.

The sentence *"Essentially, all models are wrong, but some are useful"* from the famous statistic George Box (1987) summarizes the entire study. As a matter of fact, although these predictive models were approximations of reality, it is expected that the results from this investigation will have a high impact on clinical and medical practice. In particular, predictive models, as auxiliary to clinical prognosis, are essential contributions to provide an adequate answer to parents' concerns and for a better human and technical resources allocation at neonatal care units.

**Keywords:** PREDICTIVE MODELS, LOGISTIC REGRESSION, DECISION TREES, EXTREMELY PREMATURE NEWBORNS, MORTALITY, MORBIDITY.

# Conteúdo

**Agradecimentos** iii

**Resumo** v

**Abstract** vii

**Índice de Tabelas** xi

**Índice de Figuras** xiii

**1 Introdução** 1
- 1.1 O que é o limite de viabilidade? . . . 2
- 1.2 Objetivos do trabalho . . . 3
  - 1.2.1 Metodologias utilizadas neste trabalho . . . 3
  - 1.2.2 Estrutura da tese . . . 5
- 1.3 Entidades envolvidas . . . 6
  - 1.3.1 Dados experimentais . . . 6

**2 Identificação de fatores de risco precoces** 11
- 2.1 Modelos Lineares Generalizados . . . 12
- 2.2 Regressão Logística Binária . . . 14
  - 2.2.1 Estimação dos coeficientes do modelo . . . 17
  - 2.2.2 Avaliação da qualidade do modelo . . . 18
  - 2.2.3 Interpretação do modelo . . . 21
- 2.3 Preparação da amostra e desenho do estudo . . . 23
- 2.4 Resultados e Discussão . . . 25
  - 2.4.1 Fatores de risco de mortalidade . . . 25
  - 2.4.2 Fatores de risco de morbilidade . . . 31
- 2.5 Conclusões . . . 33

**3 Modelos Preditivos** 35
- 3.1 Construção de modelos preditivos . . . 36
  - 3.1.1 Preparação da amostra . . . 37
  - 3.1.2 Número de variáveis no modelo . . . 40
  - 3.1.3 Associação entre variáveis . . . 42
  - 3.1.4 Desempenho do modelo . . . 46
- 3.2 Modelos baseados em árvores de decisão . . . 49
  - 3.2.1 Particionamento Recursivo Binário - construção da árvore . . . 51
  - 3.2.2 Critério de paragem . . . 56

3.2.3 Poda . . . 57
3.2.4 Valores em falta . . . 63
3.2.5 Modelos com custos de má classificação . . . 63
3.2.6 Resultados . . . 65
3.3 Modelos baseados em regressão logística . . . 72
3.3.1 Abordagem para procura exaustiva de modelos . . . 73
3.3.2 Resultados . . . 75
3.4 Discussão técnica dos resultados . . . 88
3.5 Conclusões . . . 90

**4 Limite de viabilidade** **91**
4.1 Discussão contextualizada dos fatores de risco . . . 92
4.1.1 Mortalidade . . . 92
4.1.2 Morbilidade . . . 97
4.2 Desempenho dos modelos preditivos e impacto na prática clínica . . . 101

**5 Conclusões** **105**

**6 Contributos científicos** **109**

**Referências** **126**

# Lista de Tabelas


1.1 Classes de desfecho (siglas e descrição). . . . . 7
1.2 Variáveis Mortality e Morbidity (codificação, onde o valor 1 representa em ambos os casos o outcome negativo). . . . . 8
1.3 Variáveis do estudo de mortalidade/morbilidade. . . . . 9

2.1 Modelos considerados para a identificação de fatores de risco precoces, onde $\pi$ representa a probabilidade de sucesso do evento (mortalidade ou morbilidade) e $\beta_j$ os coeficientes de regressão associados às respetivas variáveis $X_j$. . . . . 24
2.2 OR para a mortalidade e respetivo intervalo de confiança a 95% para os fatores de risco protocolares. . . . . 28
2.3 OR para a mortalidade e respetivo intervalo de confiança a 95% para os fatores de risco intrínsecos. . . . . 28
2.4 OR para a morbilidade e respetivo intervalo de confiança a 95% para os fatores de risco protocolares. . . . . 32
2.5 OR para a morbilidade e respetivo intervalo de confiança a 95% para os fatores de risco intrínsecos. . . . . 32

3.1 Número adequado de variáveis para modelos de regressão logística e número de observações num nó para árvores de classificação, de acordo com o tamanho da amostra. . . . . 42
3.2 Valores do coeficiente de correlação cofenético e da deformação delta de acordo com os índices do *mínimo*, *do máximo*, *da média*, *de Ward* e *do centro de gravidade* para o caso de mortalidade (preto) e morbilidade (azul), respetivamente. . . . . 44
3.3 Associação entre a variável NIV e Epoch na amostra de treino de mortalidade (preto) e morbilidade (azul), respetivamente. . . . . 46
3.4 Matriz de confusão, onde $w_1$ representa a classe do evento em estudo que pretendemos prever (Dead ou Severe) e $w_0$ a classe contrária. . . . . 47
3.5 Cálculos do decréscimo de impureza efetuados referentes aos dois exemplos ilustrados na figura 3.9. . . . . 56
3.6 Estimativas de erros de validação cruzada para a escolha da subárvore ótima referente a um ano de mortalidade. . . . . 62
3.7 Estimativas de erro de validação cruzada para a escolha da subárvore ótima referente a dois anos de morbilidade. . . . . 68
3.8 Custos a atribuir para obtenção do modelo otimista e pessimista referente a um ano de mortalidade. . . . . 70
3.9 Custos a atribuir para obtenção do modelo otimista e pessimista referente a dois anos de morbilidade. . . . . 72
3.10 Expressões dos critérios de informação AIC e BIC. . . . . 74
3.11 Tabela de classificação referente a um ano de mortalidade para um ponto de corte de 0.32, considerando o método da regressão logística. . . . . 81

3.12 Modelos de mortalidade e respetiva significância de variáveis de acordo com o teste de Wald. 82
3.13 Modelos de mortalidade após exclusão das variáveis mais promissoras de morbilidade e respetiva significância de variáveis de acordo com o teste de Wald. . . . . 83
3.14 Dez melhores modelos de mortalidade considerando toda a pesquisa exaustiva e respetiva significância de variáveis de acordo com o teste de Wald. . . . . 83
3.15 Tabela de classificação referente a dois anos de morbilidade para um ponto de corte no intervalo ]0.1379310,0.4117647[, considerando o método da regressão logística. . . . . 86
3.16 Modelos de morbilidade e respetiva significância de variáveis de acordo com o teste de Wald. 87
3.17 Dez melhores modelos de morbilidade até 3 variáveis considerando toda a pesquisa exaustiva e respetiva significância de variáveis de acordo com o teste de Wald. . . . . 87
3.18 Desempenho dos modelos de mortalidade e morbilidade em ambos os métodos estudados. . 89

4.1 Comparação dos modelos de regressão logística de mortalidade e morbilidade obtidos nesta investigação com os obtidos no estudo paralelo de Sá et al. (2012b) e Sá et al. (2012c). . . 102

# Lista de Figuras


1.1 Classificação atribuída aos recém nascidos prematuros, de acordo com o outcome. . . . . 7

2.1 Curva de regressão logística para o caso $\beta_j > 0$. . . . . 16
2.2 $OR$ para a mortalidade e respetivos intervalos de confiança a 95% para cada fator de risco protocolar (esquerda) e intrínseco (direita) utilizando regressão logística univariada (modelo Crude, Tabela 2.1). . . . . 26
2.3 Boxplots para comparação das medianas populacionais da idade gestacional e do peso de acordo com o género. Os pontos em cada caixa representam o valor médio amostral. . . . 30
2.4 Histograma do *p-value* associado à variável Multifetal Gestation, considerando *bootstrap.* . 33

3.1 Esquema ilustrativo das etapas adotadas neste trabalho para a construção de um modelo baseado em análise supervisionada. . . . . 37
3.2 Preparação e divisão da amostra inicial para a construção de modelos preditivos de (a) mortalidade e de (b) morbilidade. . . . . 38
3.3 Exemplificação do método de validação cruzada *V-fold*, adotado neste trabalho. Figura inspirada em Hastie et al. (2009). . . . . 40
3.4 Dendrogramas ilustrando o agrupamento das variáveis do estudo, baseados no índice da média e dissemelhança **1-V**, onde **V** representa o **coeficiente de Cramer V** (equação (3.4)). 45
3.5 Exemplo do espaço ROC para um classificador discreto e um contínuo. . . . . 48
3.6 Exemplo de uma árvore de classificação. Por exemplo, um bebé que apresente GA$\geq$ 25 weeks, MJD Inborn Delivery= Yes e Weight<628 g é classificado como morto. Neste trabalho adiciona-se a cada nó terminal, #**Dead** | #**Alive** da amostra de treino. . . . . 50
3.7 Principais etapas na construção de uma árvore de decisão. . . . . 51
3.8 Duas possíveis representações para a árvore máxima de mortalidade. Onde, em cada nó se observa # **Dead** | # **Alive** na amostra de treino. . . . . 52
3.9 Dois exemplos de escolha de divisão de um nó através do índice de Gini. Caso I e II respetivamente. . . . . 55
3.10 Processo de poda da árvore de classificação referente a um ano de mortalidade. O processo de poda identificou a variável Maternal age como a menos relevante na árvore de decisão. . 58
3.11 Árvore máxima $T_1$ destacando um nó candidato a ser podado. . . . . 60
3.12 Estimativas do erro de validação cruzada em função do número de nós terminais (média± SE ). . . . . 62
3.13 Árvore podada e mosaico referentes a um ano de mortalidade, onde em cada nó se observa # **Dead** | # **Alive** na amostra de treino. . . . . 66
3.14 Tabela de classificação e curva ROC referentes a um ano de mortalidade, considerando árvores de classificação. . . . . 67
3.15 Árvore máxima referente a dois anos de morbilidade, onde em cada nó se observa # **Severe** | # **Non Severe** na amostra de treino. . . . . 68
3.16 Estimativas do erro de validação cruzada em função do número de nós terminais (média±SE). 68

3.17 Árvore podada e mosaico referentes a dois anos de morbilidade, onde em cada nó se observa # **Severe** | # **Non Severe** na amostra de treino. . . . . 69
3.18 Tabela de classificação e curva ROC referentes a dois anos de morbilidade, considerando árvores de classificação. . . . . 69
3.19 Tabela de classificação e curva ROC referentes a um ano de mortalidade, para o modelo otimista (azul) e pessimista (laranja). . . . . 71
3.20 Tabela de classificação e curva ROC referentes a dois anos de morbilidade para o modelo pessimista. . . . . 72
3.21 Valor do critério BIC para cada ordem do modelo de mortalidade: (a) de acordo com a ordem do modelo (b) ordenado por ordem crescente de valor de BIC. . . . . 76
3.22 Valor do critério BIC para cada ordem do modelo de mortalidade, após exclusão das variáveis mais promissoras de morbilidade: (a) de acordo com a ordem do modelo (b) ordenado por ordem crescente de valor de BIC. . . . . 77
3.23 Curvas ROC para o estudo de um ponto de corte ótimo referente ao modelo de morbilidade. 80
3.24 Precisão global do modelo da equação (3.35) de acordo com o ponto de corte. . . . . 81
3.25 Valor do critério BIC para cada ordem do modelo de morbilidade: (a) de acordo com a ordem do modelo (b) ordenado por ordem crescente de valor de BIC. . . . . 84
3.26 Curva ROC para o estudo de um ponto de corte ótimo referente ao modelo de morbilidade. 86

4.1 Modelo preditivo de mortalidade baseado em regressão logística (esquerda) e em árvores de classificação (direita). . . . . 93
4.2 Modelo preditivo de morbilidade baseado em regressão logística (esquerda) e em árvores de classificação (direita). . . . . 98

# Capítulo 1

# Introdução

Artigos recentes, mostram que a taxa de mortalidade associada a recém nascidos prematuros extremos ($< 27$ semanas de gestação) tem vindo a decrescer, refletindo a evolução e melhoria dos cuidados médicos perinatais, bem como, das mudanças contínuas de atitude por parte médica e parental (Boussicault et al., 2012; The Express Group Members, 2010).
No entanto, a possibilidade de estes seres vulneráveis terem um desenvolvimento neurológico adverso a longo prazo é bastante elevada. Face a estas situações, o tratamento de recém nascidos prematuros levanta ainda muitas questões quer a nível médico, ético ou social (Berger et al., 2011; Wang et al., 2011). Neste sentido, a definição do *limite de viabilidade* em recém nascidos prematuros tem sido objeto de inúmeras investigações internacionais. Em Portugal, são ainda reduzidos os estudos a refletir esta problemática.

Este estudo é ainda motivado pela necessidade de atualização permanente de informação, uma vez que a viabilidade terá tendência a aumentar com o decorrer dos anos e com a evolução científica até um limite máximo que advém da limitação da capacidade intrínseca para sobrevivência (Peixoto et al., 2004). Assim sendo, esta dissertação procura refletir e caraterizar a realidade portuguesa no período de janeiro de 2000 a dezembro de 2009.

Um prognóstico mais preciso e atempado, permitiria decisões/intervenções mais adequadas, assim como um aconselhamento parental mais esclarecedor. Neste contexto, o desenvolvimento de modelos preditivos capazes de antecipar os desfechos de mortalidade/morbilidade em recém nascidos prematuros extremos são considerados de elevada importância (Medlock et al., 2011).

Neste capítulo introdutório ao trabalho da tese, será apresentado um enquadramento do *limite de viabilidade* em recém nascidos prematuros extremos de acordo com o reportado atualmente na literatura. Adicionalmente, os objetivos e as metodologias adotadas são também delineados. Por fim, é descrita a informação da base de dados conducente à elaboração de toda a investigação.

## 1.1 O que é o limite de viabilidade?

Atualmente, o grande desafio da neonatologia tem sido definir o denominado "limite de viabilidade"(Castro et al., 2011). Segundo descrito no dicionário da língua portugesa[1] "viabilidade" significa *"qualidade do que pode viver"*. Já do ponto de vista neonatal, viabilidade poderá ser definido como *"o potencial para sobreviver"* (Peixoto et al., 2004). A definição do limite de viabilidade em recém nascidos prematuros é conhecida como o limiar abaixo do qual a sobrevivência e/ou o aceitável desenvolvimento neurológico destes recém nascidos são muito improváveis. No entanto, tem sido particularmente difícil delimitar assertivamente este limite, pelo que é habitual encontrar linhas orientadoras, normalmente definidas de forma intervalar, para intervenção em prematuros extremos. Por este motivo, muitas vezes limite de viabilidade é referido como *"zona cinzenta"* (Seri and Evans, 2008).

É usual relacionar-se viabilidade com a idade gestacional de um recém nascido prematuro, talvez porque a idade gestacional é uma das variáveis mais preditivas de mortalidade e morbilidade. Estudos recentes na Suíça e em Portugal, indicam que o limite de viabilidade se refere a recém nascidos entre as 22 e as 26 semanas de gestação completas (Berger et al., 2011; Silva and Carvalho, 2008).

Portugal mantém um registo nacional e contínuo da sobrevida e das sequelas dos recém nascidos de muito baixo peso (incluiu os prematuros extremos e não só) desde 1994 (Peixoto et al., 2004). A situação nacional até 2001 indica que:

- abaixo das 23 semanas de gestação, a sobrevivência é rara e quando acontece, todos têm sequelas graves.
- às 24s+0 dias, a probabilidade de sobreviver aproxima-se dos 50%, embora apenas entre 15 a 20% sobrevivem sem sequelas.
- a partir das 25 semanas, a probabilidade de sobreviver ultrapassa os 50%, e o risco de sequelas graves é bastante baixo.

De acordo com estas evidências, aos recém nascidos com idade gestacional abaixo das 24 semanas são prestados apenas cuidados paliativos (Silva and Carvalho, 2008). No entanto, recém nascidos com idades superiores a 25 semanas devem, de forma geral, ser submetidos a reanimação imediata e tratamentos em unidades de cuidados intensivos, permanecendo a questão do que fazer aos recém nascidos entre as 24s+0 dias e as 25s+0 dias de gestação (Peixoto et al., 2004).

O limite de viabilidade também é muitas vezes estabelecido pelo peso do recém nascido, além da idade gestacional. Um artigo de revisão bibliográfica publicado recentemente (Castro et al., 2011), afirma que é consensual na literatura que crianças com idade gestacional superior ou igual a 25 semanas ou com peso maior ou igual a 600g, apresentam maturidade suficiente para sobreviver, devendo portanto ser assistidas. Contrariamente, recém nascidos com idade gestacional inferior a 24 semanas ou peso menor a 500g são considerados muito imaturos não sendo assistidos na maioria das vezes, situação equivalente ao reportado na realidade Portuguesa. No entanto, permanece a incerteza de como atuar perante recém nascidos prematuros entre as 24s+0 dias e as 25s+0 dias de gestação ou com peso entre 500 e 600 g, devendo esta decisão ser baseada em diversos fatores, quer clínicos quer familiares. Assim, o artigo publicado por Lantos and Meadow (2009) aborda diferentes formas de agir, sempre que se considerar a idade gestacional do recém nascido.

[1] Dicionário da Língua Portuguesa, 7ª Edição, Porto Editora.

Procedimentos médicos e opções tomadas por famílias perante um recém nascido prematuro são também analisados no trabalho de Kaempf et al. (2009).
Neste contexto, não é tarefa fácil caraterizar um limite de viabilidade, dado que definir o que torna a qualidade de vida aceitável a longo prazo não é um aspeto trivial.

O limite de viabilidade tem-se baseado, maioritariamente, apenas na idade gestacional ou incluindo também o peso, uma vez que estas são as variáveis de referência mais importantes. No entanto, decisões oriundas destas duas variáveis não têm permitido fazer um prognóstico adequado do desfecho do recém nascido e, portanto, não são suficientes para dar uma resposta esclarecida (Castro et al., 2011).
Estudos recentes têm procurado outros fatores que poderão dar informações úteis para definir diretrizes e intervenções médicas com o intuito de melhorar o prognóstico destes prematuros extremos (Medlock et al., 2011; Tyson et al., 2008). Adicionalmente, estes autores referem que modelos multivariados contendo alguns fatores de risco conseguem prever mais adequadamente a ocorrência de mortalidade/morbilidade do que considerar modelos que constem apenas da idade gestacional ou do peso, em separado (Medlock et al., 2011; Tyson et al., 2008).

Em concordância com o descrito anteriormente, o estudo de mortalidade e morbilidade de recém nascidos prematuros extremos ($< 27$ semanas de gestação) permanece, ainda, muito aquém de ser assertivo. Estas crianças são de facto muito imaturas, apresentam idades gestacionais muito baixas em que na maioria dos casos, a possibilidade de sobrevida e/ou sobrevida sem sequelas é pouco provável. Assim, torna-se necessário debater quando "vale a pena"investir numa determinada situação.

## 1.2 Objetivos do trabalho

O **objetivo global** desta investigação é fornecer informação útil para médicos obstetras e pediatras quanto à expectativa real da evolução destes recém nascidos prematuros extremos ($< 27$ semanas de gestação) e, assim, contribuir para um planeamento mais atempado da intervenção médica e um aconselhamento parental mais informativo.

Um primeiro **objetivo mais específico** desta investigação procura identificar e avaliar quais os fatores de risco mais precoces que influenciam a mortalidade (avaliada ao fim de um ano) e o desenvolvimento neurológico (indicador de morbilidade, acedido ao segundo ano de vida) de um recém nascido pré-termo numa Unidade Materno-Infantil em Portugal.

**Outro propósito** deste trabalho incide na construção de modelos estatísticos que permitam predizer a mortalidade e o desenvolvimento neurológico severo destas crianças, ajustado à realidade portuguesa.

### 1.2.1 Metodologias utilizadas neste trabalho

O desenvolvimento de modelos preditivos tem sido um fator crucial para a investigação em medicina clínica (Bellazzi and Zupan, 2006). A formulação do problema em estudo passou inicialmente pela decisão de que os modelos preditivos de mortalidade e morbilidade seriam construídos separadamente.

Assim, nesta investigação serão considerados dois estudos paralelos: um de mortalidade e outro de morbilidade. Esta decisão é justificada no facto de a mortalidade poder ser avaliada ao primeiro

ano de vida, enquanto que uma avaliação adequada do estado neurológico dos recém nascidos prematuros extremos só é possível bastante tempo depois do nascimento (segundo ano de vida), o que implica conduzir estudos de *follow up* mais prolongados no tempo para obter dados adequados aos objetivos. Desta forma, considerando modelos separados, ter-se-á a vantagem de poderem ser selecionadas variáveis mais apropriadas a cada desfecho. Posteriormente, com base numa pesquisa bibliográfica, estudos de modelos preditivos em condições equivalentes, indicam que considerar modelos combinados de mortalidade/morbilidade muitas vezes não são os mais adequados, uma vez que são dois outcomes distintos e as possíveis variáveis preditoras podem não ser exatamente as mesmas (Ambalavanan et al., 2006). Além do mais, modelos nestas condições mostram ter um desempenho inferior quando comparados com modelos separados (Medlock et al., 2011).

Para cada estudo, serão construídos modelos baseados em dois métodos distintos: regressão logística e árvores de decisão.

Por um lado, a regressão logística é, segundo um artigo de revisão recentemente publicado por Medlock et al. (2011), o método mais utilizado na previsão do desfecho de recém nascidos prematuros extremos. Adicionalmente, esta é aplicada constantemente para identificação de fatores de risco em vasta literatura clínica, uma vez que consegue enfatizar qual a gravidade de tal risco (Bagley et al., 2001). Desta forma, é objetivo deste trabalho construir modelos que sejam comparáveis com os reportados na literatura médica para a realidade de outros países. O método de regressão logística permite estimar a probabilidade de um determinado recém nascido prematuro estar associado a mortalidade/morbilidade, bem como determinar a significância estatística das variáveis na previsão dos outcomes.

Por outro lado, há interesse em construir um modelo adequado e intuitivo para os especialistas na área clínica. Neste sentido, as árvores de decisão foram, sem dúvida, um método complementar ao método de regressão logística devido à sua estrutura lógica e consequente facilidade de interpretação. Estas constituem um dos métodos de classificação mais intuitivos, dado que apresentam uma estrutura hierárquica meramente baseada num conjunto de regras.

Adicionalmente, as árvores têm ainda a vantagem de conseguirem lidar com valores em falta, situações muito frequentes em bases de dados na área da saúde e que se revêm no presente caso em estudo. Os valores omissos podem ocorrer pela questão de os profissionais de saúde não terem registado o dado correspondente, pelo paciente deixar de ser seguido no hospital em questão ou mesmo devido à ocorrência de morte no período de recolha do dado. Determinar a abordagem adequada para este tipo de conjunto de dados pode tornar-se um processo bastante delicado, na medida em que a escolha de métodos não apropriados pode levar a conclusões erradas sobre a população. Recentemente, têm sido estudadas metodologias capazes de substituir os valores em falta de acordo com toda a informação presente na base de dados (Antunes et al., 2011). Não obstante, neste trabalho de investigação optou-se por não se utilizar qualquer método de imputação de dados, uma vez que grande parte dos valores em falta presentes referem-se a casos em que os dados não puderam ser recolhidos porque os recém nascidos morreram precocemente. Neste contexto, os valores em falta serão excluídos/incluídos consoante o método adotado. Em alternativa, poder-se-ia ter optado por utilizar apenas o método de regressão logística fazendo imputação de tais valores.

Pelos motivos já descritos, as árvores de decisão estão a abranger cada vez mais a área médica, embora em menor expressão quando comparadas com a regressão logística.
Podgorelec et al. (2002) descrevem detalhadamente inúmeras aplicações deste método em diversos

campos da medicina. Problemas referentes a fraturas ortopédicas, diagnósticos de enfarte de miocárdio e fatores que indiciem um parto por cesariana são apenas alguns dos exemplos de aplicação nele citados. Uma aplicação de árvores de decisão inserida no tema desta investigação é também abordada no trabalho de Ambalavanan et al. (2006).
Segundo Breiman et al. (1984), as árvores de decisão podem ser de dois tipos: árvores de classificação (do inglês, *classification tree*), quando a variável a prever é uma variável nominal ou árvores de regressão (do inglês, *regression tree*), quando a variável a prever é uma variável quantitativa. Neste trabalho, serão apenas abordadas as árvores de classificação, uma vez que as variáveis de mortalidade/morbilidade a predizer são nominais (subsecção 1.3.1).

A aplicação das metodologias adotadas foi implementada com recurso ao *software R*, versão 2.15.0 (R Development Core Team, 2012). Adicionalmente, foram também desenvolvidas rotinas de visualização gráfica para uma melhor perceção dos resultados obtidos.

### 1.2.2 Estrutura da tese

Este trabalho de investigação encontra-se segmentado em diversos capítulos, onde a primeira página apresenta um resumo do que será tratado nesse mesmo capítulo e na última página, uma conclusão sucinta sobre os resultados obtidos. Esta organização permite ao leitor ter uma ideia geral da investigação, remetendo os detalhes do estudo para o corpo integral do capítulo.

A introdução ao tema e o objetivo são apresentados no **capítulo 1**. É efetuada uma contextualização de caráter introdutório relativamente a recém nascidos prematuros no *limite de viabilidade* bem como o seu enquadramento no meio matemático. Adicionalmente, são definidos os objetivos desta investigação.

O **capítulo 2** inclui a base teórica que sustenta todo o processo de identificação de fatores de risco. Posteriormente, procura-se clarificar o estado atual da população de recém nascidos prematuros extremos relativamente a fatores de risco precoces considerados significativos de mortalidade/morbilidade.

O **capítulo 3** inicia com uma abordagem de índole teórica sobre modelos preditivos. São destacadas algumas das etapas importantes na sua construção, tanto para metodologias de árvores de decisão como de regressão logística. São também consideradas abordagens alternativas aos métodos de seleção de variáveis mais comuns, nomeadamente agregação entre variáveis e pesquisa exaustiva de modelos considerando um limiar de variáveis a incluir. Os resultados para mortalidade e morbilidade de recém nascidos prematuros extremos são apresentados e comparados tecnicamente.

A análise crítica e contextualizada dos resultados é apresentada no **capítulo 4**. Em particular, os resultados obtidos no capítulo 2 e 3 são discutidos do ponto de vista clínico e comparados com estudos de referência na literatura. Por fim, os modelos são ainda avaliados quanto ao seu desempenho, repercussão e impacto na prática clínica.

No **capítulo 5** são apresentadas as principais conclusões, as limitações e considerações finais relevantes, bem como algumas recomendações para trabalho futuro.

Por fim, no **capítulo 6** são listados os contributos científicos desenvolvidos no âmbito deste trabalho e já apresentados em conferências/congressos da especialidade de estatística e obstetrícia/ neonatologia.

## 1.3 Entidades envolvidas

Este trabalho de investigação surge no âmbito do Mestrado em Engenharia Matemática (engmat, *http://www.fc.up.pt/dmat/engmat/*), lecionado no Departamento de Matemática da Faculdade de Ciências da Universidade do Porto (DMAT/FCUP). A realização do mesmo resulta de uma parceria entre o Gabinete de Estatística, Modelação e Aplicações Computacionais (GEMAC, *http://cmup.fc.up.pt/cmup/gemac/*), o Centro de Matemática da Universidade do Porto (CMUP, *http://cmup.fc.up.pt/cmup/*) e a Unidade Maternidade de Júlio Dinis do Centro Hospitalar do Porto (MJD-CHP, *http://www.chporto.pt/ver.php?cod=0A0C0C*). Desta última instituição, colaboraram neste trabalho, especialistas dos serviços de Neonatologia e Ginecologia/Obstetrícia.

O GEMAC é um gabinete de consultadoria integrado no CMUP cujo principal objetivo é promover parcerias em projetos de Investigação e Desenvolvimento. Este é constituído essencialmente por especialistas na área da estatística.
O CMUP refere-se a um centro de investigação cujo objetivo primordial é, desde a sua fundação, apoiar a investigação Matemática e promover a divulgação desta como uma disciplina fundamental para o desenvolvimento da sociedade. Está integrado no DMAT/FCUP e dele fazem parte diversos investigadores da área.
A Maternidade de Júlio Dinis (MJD) trata-se de um Hospital Central Especializado, incluído na rede nacional de hospitais do Serviço Nacional de Saúde, prestando cuidados na área da Mulher e da Criança. Situada na cidade do Porto e inaugurada em Setembro de 1939, a MJD é considerada uma referência na assistência Materno Infantil, primando pelo prestígio profissional dos especialistas em saúde que lá trabalham. Desde 2007, a MJD foi integrada no Centro Hospitalar do Porto (CHP), designando-se atualmente de CHP-Unidade Maternidade de Júlio Dinis.
A equipa de investigação referente a este trabalho é de facto multidisciplinar, incluindo um grupo de profissionais de saúde da MJD. Desta forma, no contexto desta pesquisa procedeu-se uma investigação conjunta em que parte dela se encontra reportada nesta tese e a restante, mais vocacionada para a parte médica é utilizada neste estudo como comparação.

### 1.3.1 Dados experimentais

A presente tese incide sobre dados anónimos de 205 recém nascidos prematuros extremos (<27 semanas de gestação) seguidos na MJD entre janeiro de 2000 e dezembro de 2009 e resulta de uma colaboração estreita entre os serviços de Ginecologia/Obstetrícia e Neonatologia da MJD com aprovação pela Comissão de Ética do CHP. Durante este período de 10 anos, ocorreram 33 638 partos na MJD, dos quais 183 ocorreram antes das 27 semanas de gestação. No mesmo período, houveram 4 245 admissões na unidade de Neonatologia, sendo 169 dos recém nascidos idades gestacionais inferiores a 27 semanas de gestação. A base de dados cedida por esta unidade materno-infantil portuguesa fornece o outcome destes recém nascidos prematuros de acordo com uma divisão em classes, consoante a ocorrência de mortalidade ao fim de um ano e/ou morbilidade ao segundo ano (momento em que se avalia a existência de sequelas major).
A figura 1.1 ilustra as várias classes adotadas, sendo que esta divisão é frequentemente utilizada em estudos semelhantes noutros países (The Express Group Members, 2010). A tabela 1.1 apresenta a designação de cada classe bem como a representatividade das mesmas na base de dados.

Neste trabalho de investigação, as duas primeiras classes, SB e DRD foram excluídas do estudo, à semelhança dos trabalhos de Kaiser et al. (2004), Medlock et al. (2011), The Express Group Members (2010) e outros. Esta exclusão é justificada pelo facto de estes recém nascidos não terem tempo para usufruir da maioria dos cuidados médicos prestados, uma vez que ou já nascem mortos

ou morrem na sala de parto.

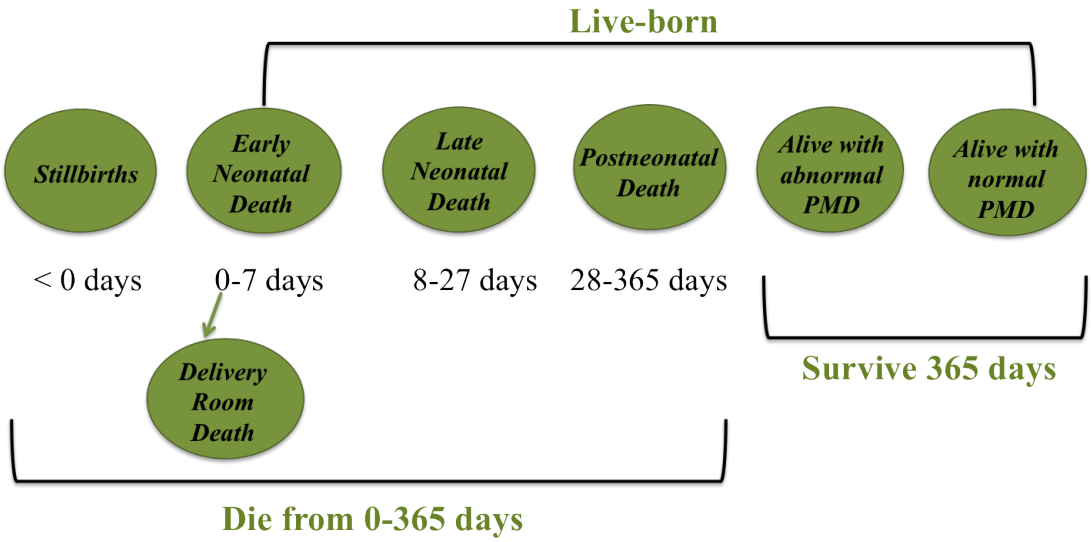



Figura 1.1: Classificação atribuída aos recém nascidos prematuros, de acordo com o outcome.

Tabela 1.1: Classes de desfecho (siglas e descrição).

| **Sigla** | **Classe** | **Designação** | **Frequência #(%)** | **Frequência #(%)** |
|---|---|---|---|---|
| SB | Stillbirths | Nados mortos | 30(14.6%) | |
| DRD | Delivery Room Death | Morte na sala de parto | 6(2.9%) | 36 (17.5%) |
| END | Early Neonatal Death | Morte Precoce | 67(32.7%) | |
| LND | Late Neonatal Death | Morte Tardia | 18(8.8%) | 89 (43.4%) |
| PND | Postneonatal Death | Morte Pós neonatal | 4(1.9%) | |
| AAbn | Alive with abnormal PMD* | Sobrevive com DPM⋆ anornal | 15(7.3%) | 74 (36.1%) |
| AN | Alive with normal PMD* | Sobrevive com DPM⋆ normal | 59(28.8%) | |
| – – – | Missing Data | – – – | 6 (2.9%) | 6 (2.9%) |
| | **Total** | | | 205(100%) |

*designa psycomotor development e ⋆designa desenvolvimento psicomotor

Adicionalmente, constata-se que não há tamanho amostral para se realizar um estudo de mortalidade/morbilidade pormenorizado em classes, dado que o tamanho da amostra em cada classe não permite fazer um estudo estatístico *válido* (Tabela 1.1). Inerente a esta limitação ao tamanho da amostra e de forma a que o estudo possa ser viável, optou-se por mergir classes. Porém, esta estratégia foi adotada supondo homogeneidade entre as mesmas. Neste contexto, foram criadas duas variáveis:

- Mortality
- Morbidity

resultantes da reagrupação das classes iniciais e que posteriormente foram codificadas de forma binária. A tabela 1.2 sumaria as variáveis Mortality/Morbidity e respetiva codificação, as quais serão consideradas como outcomes de mortalidade/morbilidade ao longo deste trabalho. Note-se que o valor 1 representa, em ambos os casos, o outcome negativo.

No caso da variável Mortality, o valor 1 corresponde a Dead (morto), enquanto que o valor 0 corresponde a Alive (vivo). Estas duas novas classes referem-se às uniões:

- Dead= END $\cup$ LND $\cup$ PND;
- Alive= AAbn $\cup$ AN.

Já para a morbilidade, o valor 1 corresponde a Severe (morbilidade severa), sendo que 0 traduz os recém nascidos com morbilidade não severa. As classes Severe e Non Severe correspondem a:

- Severe= AAbn;
- Non Severe= AN.

Saliente-se que as observações da morbilidade referem-se a um subconjunto de mortalidade correspondente aos prematuros extremos que sobreviveram ao fim de um ano (classe Alive da variável Mortality), como sugere a figura 1.1. É também importante realçar que enquanto a variável Mortality é acedida ao fim de um ano, a avaliação do estado neurológico da criança (variável Morbidity) só é efetuada aos dois anos de idade.

Tabela 1.2: Variáveis Mortality e Morbidity (codificação, onde o valor 1 representa em ambos os casos o outcome negativo).

| **Variável** | **Descrição** | **Codificação** |
|---|---|---|
| Mortality | Mortalidade | 1:Dead; 0:Alive |
| Morbidity | Morbilidade | 1:Severe; 0:Non Severe |

Adicionalmente, a base de dados é também constituída por diversas variáveis respeitantes à mãe e ao recém nascido, tais como dados maternos e história obstétrica, detalhes sobre a gravidez e o parto, condições do recém nascido prematuro extremo no nascimento e procedimentos neonatais. Foram ainda efetuadas algumas alterações na codificação das variáveis com o propósito de se manter uma certa coerência na codificação e assim facilitar a interpretação dos resultados. Em particular, todas as variáveis binárias foram recodificadas, assumindo-se sempre o valor 1 como "sim"e 0 como "não". A tabela 1.3 apresenta tais variáveis que serão incluídas no estudo de mortalidade/morbilidade e respetiva codificação.

Como se pode constatar, a maioria das variáveis em estudo são categóricas: dicotómicas ou politómicas. Devido às suas caraterísticas, estas variáveis categóricas não serão consideradas como ordinais em nenhuma etapa desta dissertação, uma vez que a diferença entre se considerar ou não ordenação verificou ser de importância menor neste trabalho.

Tabela 1.3: Variáveis do estudo de mortalidade/morbilidade.

| **Variável** | **Acrónimo** | **Legenda** | **Codificação** |
|---|---|---|---|
| Weight (g) | − − − | Peso (g) | escala (gramas) |
| Gender | − − − | Género | 1:masculino; 0:feminino |
| Maternal age age (years) | − − − | Idade da Mãe (anos) | escala (anos) |
| Pregnancy Surveillance | PS | Vigilância | 1:sim; 0:não |
| Primipara | − − − | Primípara | 1:sim; 0:não |
| Multifetal Gestation | − − − | Gestação múltipla | 1:sim; 0:não |
| Antenatal Steroids | − − − | Corticóides | 1:sim; 0:não |
| Membrane Rupture | MR | Rutura de membrana | 1:<12h; 2:12-24h; 3:≥24h |
| GA(weeks) | − − − | Idade gestacional (semanas) | escala |
| Endotracheal/ ETT resuscitation | ETT resuscitation | Necessidade de intubação | 1:sim; 0:não |
| Bronchopulmunary Dysplasia | BPD | Displasia broncopulmonar | 1:sim; 0:não |
| Caesarean Delivery | − − − | Parto por cesariana | 1:sim; 0:não |
| Iatrogenic Delivery | − − − | Parto iatrogénico | 1:sim; 0:não |
| Epoch | − − − | Época | 1:(2000-2002); 2:(2003-2006); 3:(2007-2009) |
| Intraventricular Hemorrage | IVH | Hemorragia Intraventricular | 1:no; 2:grau I ou II; 3:grau III ou IV |
| Periventricular Leukomalacia | PVL | Leucomalácia Periventricular | 1:sim; 0:não |
| Small for gestational age | SGA | Pequeno para a idade gestacional | 1:sim; 0:não |
| 5-min Apgar≤3 | − − − | Apgar ≤ a 3 aos 5 minutos | 1:sim; 0:não |
| Patent Ductus Arteriosus | PDA | Persistência do canal arterial | 1:sim; 0:não |
| Surfactant | − − − | Surfactante | 1:sim; 0:não |
| Hyaline Membrane disease | HMD | Doença da membrana hialina | 1:sim; 0:não |
| Infection | − − − | Infeção | 1:sim; 0:não |
| MJD Inborn Delivery | − − − | Nascimento na MJD | 1:sim; 0:outras instituições |
| Oxygen | $O_2$ | Oxigénio | 1:sim; 0:não |
| Non invasise ventilation | NIV | Ventilação não invasiva | 1:sim; 0:não |
| Mechanical ventilation | MV | Ventilação mecânica | 1:sim; 0:não |

Embora a maioria das variáveis seja de interpretação imediata, apresenta-se de seguida uma breve descrição de algumas.

- **Índice de apgar (5-min Apgar $\leq$3)**

O índice de apgar trata-se de um procedimento clínico que consiste na avaliação de 5 sinais objetivos (frequência cardíaca, respiração, tónus muscular, irritabilidade reflexa e cor da pele) do recém nascido no primeiro, no quinto e no décimo minuto após o nascimento. O índice de apgar varia entre 0 e 10, correspondendo a atribuir no máximo 2 pontos a cada sinal. O teste é repetido 3 vezes para que seja possível prestar cuidados adequados relativamente aos sinais que mostraram "fraqueza", sendo o valor dos 5 minutos considerado o mais importante. Geralmente, índices de apgar inferiores a 7 nesta etapa indicam que o recém nascido necessita de observação.

- **Rutura de Membrana (MR)**

Denomina-se de rutura de membrana à abertura da bolsa onde se encontra o feto. Esta variável indica de forma intervalar o tempo em que a grávida permaneceu com a membrana rota antes do parto.

- **Necessidade de Intubação (ETT Resuscitaion)**

A variável ETT Resuscitation não indica que o recém nascido prematuro tenha feito intubação mas sim se este teve ou não necessidade de usufruir de tal tratamento.

- **Época (Epoch)**

Refere-se aos anos em que os dados foram recolhidos e encontra-se codificada em intervalos, de acordo com a mudança de protocolos na prática clínica da MJD. Desta forma, será possível compreender se determinados acontecimentos foram causados por tais mudanças.

- **Pequeno para a idade gestacional (SGA)**

São considerados pequenos para a idade gestacional, recém nascidos cujo peso seja 2 vezes inferior ao desvio padrão da média para cada idade gestacional, em concordância com The Express Group Members (2010).

- **Nascimento da MJD (MJD Inborn Delivery)**

O valor 1 desta variável representa os prematuros que nasceram na MJD. No entanto, o valor 0 é atribuido a recém nascidos prematuros que nasceram noutras instituições, nomeadamente em hospitais considerados menos especializados, pelo que necessitarem de ser transportados para a MJD após o nascimento e não *in-utero*.

- **BPD, IVH, PVL, HMD**

Estas variáveis são exemplos de complicações habituais de carácter respiratório, motor, cognitivo e cardiovascular, neste tipo de crianças.

- $O_2$**, NIV, MV**

Trata-se de cuidados médicos aplicáveis quando a criança se encontra com problemas respiratórios significativos, sendo prestados à grande maioria dos recém nascidos prematuros extremos.

# Capítulo 2

# Identificação de fatores de risco precoces

Neste capítulo pretende-se determinar fatores de risco precoces associados a mortalidade e morbilidade em recém nascidos prematuros extremos. A vantagem da análise se centrar em preditores precoces possibilita aos peritos em saúde diagnosticar antecipadamente diversas complicações, permitindo imediatamente tomarem decisões referentes a intervenções médicas e não menos importante, informarem/aconselharem adequadamente os pais destes recém nascidos (Boussicault et al., 2012; The Express Group Members, 2010). Os dois estudos serão realizados separadamente: um de mortalidade e outro de morbilidade, por forma a que seja possível aceder-se aos fatores de risco mais precoces correspondentes a cada desfecho separadamente. Além disto, a identificação de fatores de risco precoces visa ser um estudo preliminar à construção de modelos preditivos.

Utilizar-se-á o método de regressão logística, uma vez que este é o mais usual para determinar fatores de risco, sendo aplicado em diversas áreas mas sobretudo em medicina. Uma referência notável acerca deste método é o livro publicado por Hosmer and Lemeshow (2000), intitulado *"Applied Logistic Regression"*.

Neste capítulo a regressão logística será inicialmente introduzida como um Modelo Linear Generalizado (MLG). Em particular será dado ênfase às caraterísticas gerais dos MLG.
Posteriormente, é abordada a regressão logística binária num contexto teórico onde serão descritos alguns detalhes essenciais. Por fim, são apresentados e discutidos os resultados alcançados, fazendo-se um contraponto com o que tem vindo a ser referenciado na literatura.

Parte deste trabalho foi apresentado no 5º Encontro de Investigação Jovem da Universidade do Porto (IJUP'12) e nas XIX Jornadas de Classificação e Análise de Dados (JOCLAD2012) (Januário et al., 2012d,a).

## 2.1 Modelos Lineares Generalizados

Um modelo matemático é descrito como sendo uma interpretação/representação simplificada da realidade. Estes modelos podem ser denominados determinísticos, caso os resultados inerentes sejam definidos com precisão e, probabilísticos, quando os resultados assumem variabilidade devido a fatores aleatórios desconhecidos. Assim sendo, um modelo *estatístico* é um modelo que possui uma componente probabilística (Lindsey, 1997).

Nos últimos tempos, a *análise de regressão* tem-se tornado uma das ferramentas estatísticas mais utilizadas para a análise de dados, na medida em que esta se carateriza por ser um método conceitual capaz de traduzir uma relação funcional entre variáveis (Chatterjee et al., 2000). Esta relação é expressa na forma de uma equação, cujo pressuposto visa descrever a relação entre uma variável resposta (variável dependente) e as variáveis explicativas (variáveis independentes ou covariáveis) que descrevem o problema em estudo.
O modelo de regressão mais conhecido é a regressão linear clássica, em que a variável resposta provém de uma distribuição normal e a relação entre esta e as variáveis explicativas assume uma estrutura linear. Todavia, devido às caraterísticas dos dados estas considerações nem sempre são possíveis.

Neste contexto, Nelder and Wedderburn (1972) propuseram os *Modelos Lineares Generalizados* (MLG), caraterizados como sendo uma extensão do modelo de regressão linear clássico para respostas não normais, categóricas ou ordinais como por exemplo: dados de contagem, proporções, entre outras. A ideia base demonstrada por estes autores visa considerar os MLG como uma classe de modelos cuja distribuição da variável resposta pertence a uma única família de distribuições, *família exponencial* e que apresentam uma estrutura linear nas variáveis explicativas. A estimação dos coeficientes destes modelos é efetuada através do método da máxima verosimilhança, sendo a maximização desta função de verosimilhança obtida através de métodos numéricos iterativos.

Devido à sua vasta extensão, os MLG têm vindo a ser cada vez mais utilizados e desenvolvidos a nível de software. Existem diversas distribuições pertencentes à família exponencial e consequentemente diversos tipos de MLG que são aplicados de acordo com os dados em estudo. No que concerne ao modelo de regressão logística este é aplicado quando a resposta assume natureza dicotómica. Este tipo de regressão é o mais utilizado na área da medicina (Bagley et al., 2001), ao pretender-se por exemplo, identificar se um determinado indivíduo sobrevive ou não de acordo com vários fatores, se um determinado tratamento é ou não eficaz, ou mesmo se um indivíduo apresenta ou não uma determinada doença.

**Família Exponencial**
A variável aleatória $Y$ tem uma distribuição pertencente à família exponencial se a sua função de massa/densidade de probabilidade (caso $Y$ seja discreta/contínua respetivamente) possa ser escrita na seguinte forma:

$$f_Y(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.1}$$

para algumas funções específicas $a(\phi)$, $b(\theta)$ e $c(y,\phi)$ (Faraway, 2006; McCullagh and Nelder, 1989). $\theta$ assume-se como um parâmetro de localização, sendo designado por parâmetro canónico, e, $\phi$ denomina-se por parâmetro de dispersão, representando uma escala. De acordo com a equação (2.1) diz-se que a distribuição de $Y$ está na forma canónica.
Segundo a prova de McCullagh and Nelder (1989), se $Y$ é uma variável aleatória com distribuição

pertencente à família exponencial, em concordância com as condições anteriores, então:

$$E(Y) = \mu = b'(\theta) \quad e \quad Var(Y) = \sigma^2 = b''(\theta)a(\phi) \tag{2.2}$$

A média surge como uma função de $\theta$ e a variância como um produto de duas funções que dependem da localização e do parâmetro de escala respetivamente.

A família exponencial abrange inúmeras distribuições, por exemplo distribuições discretas tais como Binomial e Poisson ou distribuições contínuas como a distribuição Normal, Gamma ou Inversa Gaussiana. Exemplos referentes a distribuições pertencentes à família exponencial e respetivos detalhes podem ser consultados em Faraway (2006); Lindsey (1997); McCullagh and Nelder (1989); Turkman and Silva (2000).

Considere-se o exemplo seguinte que retrata o tipo de distribuição que será utilizada posteriormente neste trabalho.

*Distribuição Binomial:* Se $Y \sim B(n, \pi)$, onde $n$ é o número de experiências de um determinado evento e $\pi$ a probabilidade de sucesso desse memso evento, então a sua função de probabilidade pode ser escrita de acordo com a equação (2.1) do seguinte modo:

$$f_Y(y; \theta, \phi) = \binom{n}{y} \pi^y (1-\pi)^{n-y} =$$

$$= exp\left\{ y \log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y} \right\} \tag{2.3}$$

fazendo $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, $a(\phi) = 1$, $b(\theta) = n\log(1 + exp(\theta))$ e $c(y, \phi) = \log\binom{n}{y}$ escrevemos a equação (2.3) na forma da equação (2.1). Nestas condições, $E(Y) = b'(\theta) = n\pi$ $Var(Y) = a(\phi)b''(\theta) = n\pi(1-\pi)$.

**Componentes dos MLG**
Os MLG são caracterizados por três componentes: *componente aleatória*, *componente sistemática* e *função de ligação* (Agresti, 1996; McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972).

1. *Componente aleatória*, onde é identificada a variável resposta $Y$ e a sua respetiva distribuição de probabilidade;

   Assim, temos que:

   $$E(Y) = \mu = b'(\theta)$$

2. *Componente sistemática*, onde é explicitada a relação entre as covariáveis do modelo através de um preditor linear;

   Designem-se por $X_1, ..., X_p$ as $p$ variáveis explicativas e $\eta$ o preditor linear:

   $$\eta = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

onde $\beta = (\beta_0, ..., \beta_p)$ é o vetor dos coeficientes de regressão associados à constante e às $p$ variáveis explicativas do modelo. A expressão anterior pode ser escrita na forma matricial de acordo com:

$$\eta = X\beta$$

onde $X$ denota a matriz das $p$ variáveis explicativas e $\beta$ o vetor dos $p+1$ coeficientes de regressão do modelo. É de realçar que a primeira coluna da matriz $X$ é constituída por uns, para que se consiga para cada observação, $i = 1, ..., n$ obter-se o parâmetro constante $\beta_0$.

3. *Função de ligação*, onde se descreve a relação funcional existente entre a *componente sistemática* e o valor esperado da *componente aleatória*.

   Como o nome sugere, esta função estabelece uma ligação entre as *componentes aleatória* e *sistemática*, isto é, relaciona o valor esperado $E(Y) = \mu$ com o preditor linear $\eta$. No caso dos MLG tal ligação é estabelecida através de uma função $g(\mu)$ monótona e diferenciável, designada por *função de ligação*:

   $$g(\mu) = X\beta \tag{2.4}$$

   Neste seguimento tem-se que:

   $$\eta = g(\mu) = X\beta \tag{2.5}$$

   Note-se que a função de ligação é utilizada com o objetivo de efetuar uma transformação do $E(Y) = \mu$ para uma escala em que este não fique restrito. A função de ligação mais simples é obtida quando $g(\mu) = \mu$. Nestes casos diz-se que a função de ligação é a identidade, como é o caso do modelo de regressão linear clássico. No caso da regressão logística a função de ligação mais usual denomina-se de *logit*, como será explicitado posteriormente.

## 2.2 Regressão Logística Binária

A regressão logística difere da regressão linear, na medida em que assume que a variável resposta é *dicotómica* ou *politómica*, isto é, possui dois ou mais valores possíveis de resposta, podendo ser *nominal* ou *ordinal*. Designa-se de regressão logística *ordinal* quando existe uma ordem subjacente entre as categorias da variável resposta e regressão logística *nominal* quando no contexto do problema não existe esta ordem. Relativamente às variáveis explicativas, estas podem ser categóricas ou contínuas. Nesta secção apenas será abordada a regressão logística binária (apenas duas categorias), que constitui segundo Agresti (1996) e Hosmer and Lemeshow (2000) o método mais usual para lidar com dados binários. Quando o modelo de regressão logística é expresso por diversas variáveis explicativas, este denomina-se de *regressão logística multivariada*. Um caso particular deste é então, a chamada *regressão logística univariada*, onde existe apenas uma variável explicativa.

Seja $Y = \{0, 1\}$ a variável resposta binária, onde 1 é traduzido como o "sucesso"do evento em estudo e seja $Y_1, ..., Y_n$ uma amostra aleatória. Seja um vetor das $p$ covariáveis $X = (X_1, ..., X_p)$ e uma determinada observação $x_i = (x_{1i}, ..., x_{pi})$ do objeto $i$ assume-se que:

$$Y|(X = x_i) \sim Bin(1, \pi_i) \tag{2.6}$$

sendo $\pi_i = \pi(x_i) = P(Y = 1|X = x_i)$ a probabilidade de sucesso dado que se observou $X = x_i$ (Agresti, 1996). Analogamente, a probabilidade de fracasso é definida como sendo complementar

da anterior, $1 - \pi_i = P(Y = 0|X = x_i)$. É possível verificar, pelas propriedades do valor esperado e da variância (equação 2.2), que:

$$\begin{aligned} \mu_i = \quad E(Y_i) \quad &= E(Y|X = x_i) = \pi_i \\ \sigma^2 = \quad Var(Y_i) \quad &= Var(Y|X = x_i) = \pi_i(1 - \pi_i) \end{aligned} \tag{2.7}$$

observando-se que a média e a variância dependem ambas da probabilidade $\pi_i$ (Rodríguez, 2007). Isto sugere que modelos onde se assume variância constante, como o caso dos modelos lineares, não são apropriados para modelar dados binários.

Em problemas de regressão, a quantidade chave que se pretende modelar é $E(Y|X = x_i)$, isto é, o valor esperado da variável resposta dada uma determinada observação $x_i$ (Hosmer and Lemeshow, 2000). Nos modelos lineares clássicos este valor esperado é expresso diretamente através de uma combinação linear entre as variáveis explicativas, como expresso na equação seguinte:

$$\begin{aligned} E(Y|X = x_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} \qquad i = 1, ..., n \quad j = 1, ..., p \end{aligned} \tag{2.8}$$

onde $\beta_j$ representa os coeficientes de regressão associados às $p$ variáveis explicativas e $x_{ji}$ o valor de cada variável explicativa referente a um determinado objeto $i$. No caso particular da regressão logística, temos que $E(Y|X = x_i) = \pi_i$, e portanto a equação anterior resume-se a:

$$\pi_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} \tag{2.9}$$

É possível constatar que $0 \leq \pi_i \leq 1$ e o termo do lado direito poderá assumir valores entre $-\infty$ e $+\infty$, não garantindo que os valores previstos estejam contidos no intervalo correto.
Neste seguimento, e de acordo com a estrutura dos MLG descritos na secção 2.1 torna-se necessária a aplicação de uma função de ligação $g(\mu_i)$ apropriada a este tipo de dados que seja capaz de relacionar o valor esperado da variável resposta, $\pi_i$, com o preditor linear $\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}$. Pretendemos então que:

$$\eta_i = g(\mu_i) = g(\pi_i) \tag{2.10}$$

No caso da regressão logística, a função de ligação mais comum é o *logit* (Agresti, 1996; Hosmer and Lemeshow, 2000; McCullagh and Nelder, 1989), definida por

$$g(\pi_i) = logit(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right). \tag{2.11}$$

Após a aplicação desta transformação à equação (2.9) vem que

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}. \tag{2.12}$$

Note-se que o logaritmo faz com que o termo do lado esquerdo da equação varie entre $-\infty$ e $+\infty$ como o outro membro. Através das equações (2.6) e (2.12) temos então definido o denominado *modelo logit*.

Adicionalmente, resolvendo a equação (2.12) em ordem a $\pi_i$, obtém-se uma expressão para a probabilidade *a posteriori* de um certo objeto pertencer ao evento em estudo

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ji} \quad &\Leftrightarrow \quad \frac{\pi_i}{1-\pi_i} = e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ji}} \Leftrightarrow \\ &\Leftrightarrow \quad \pi_i = e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ji}}(1-\pi_i) \Leftrightarrow \\ &\Leftrightarrow \quad \pi_i = \frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ji}}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ji}}} \end{aligned} \tag{2.13}$$

A regressão logística permite assim estimar a probabilidade de ocorrência de um evento, $\pi_i = P(Y = 1|X = x_i)$, tendo em conta as diversas covariáveis. Para determinar o valor da variável resposta $Y_i$ (0 ou 1) associado e um determinado objeto, é usual assumir-se um valor de corte para $\pi_i$, por exemplo 0.5. Assim sendo, assume-se que se $\pi_i \geq 0.5 \rightarrow$ Y=1 (o objeto pertence à classe definida como o "sucesso"do evento), caso contrário Y=0. Um valor de corte ótimo poderá ser estimado através dos dados, por forma a aumentar o ajuste do modelo de regressão logística aos dados.

A figura 2.1 apresenta para o caso univariado, $p = 1$, o gráfico produzido pela equação (2.13). Observa-se que o parâmetro $\beta_j$ determina a forma que a curva logística assume: por um lado, o sinal de $\beta_j$ indica se a curva cresce ($\beta_j > 0$) ou decresce ($\beta_j < 0$) e por outro, o valor de $\beta_j$ em módulo indica o quão rápido esta curva cresce ou decresce (Agresti, 1996).

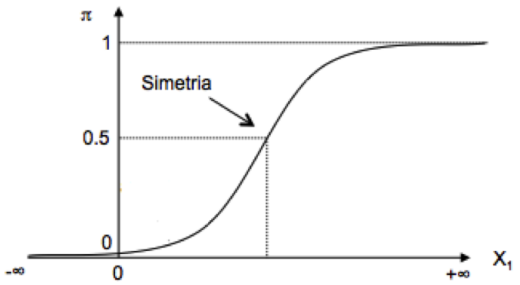



Figura 2.1: Curva de regressão logística para o caso $\beta_j > 0$.

Por fim, é importante realçar que o cálculo de $\pi_i$ não está restrito ao tipo de variáveis. Como referido no início desta secção, as variáveis explicativas, $X_1, ..., X_p$ podem ser de natureza contínua ou categórica. Caso as variáveis sejam categóricas, são então transformadas nas denominadas variáveis "dummys"(tomando apenas dois valores possíveis, 0 e 1), onde para $k$ valores possíveis de uma variável categórica são criadas $k-1$ dummys, ficando uma das categorias da variável como referência.
Nestes casos, a equação (2.12) pode ser reescrita como :

$$logit(\pi_i) = \beta_0 + \beta_1 x_{1i} + ... + \sum_{l=1}^{k_j-1}\beta_{jl}x_{jli} + \beta_p x_{pi} \tag{2.14}$$

considerando-se a j-ésima variável categórica com $j$ categorias, $x_{jli}$ representa cada variável dummy criada para o objeto $i$ e $\beta_{jl}$ os respetivos coeficientes.

### 2.2.1 Estimação dos coeficientes do modelo

Os coeficientes do modelo de regressão logística são estimados através do método da máxima verosimilhança, à semelhança dos restantes MLG (Agresti, 1996; Dobson, 2002). Pelo princípio da máxima verosimilhança, as estimativas de $\beta$, $\widehat{\beta}$, serão aquelas que maximizem a função de verosimilhança. Assumindo independência nas observações, esta função corresponde a considerar o produtório de uma função de probabilidade, $f(y|x_i)$:

$$l(\beta, y) = \prod_{i=1}^{n} f(y|x_i) \, , \tag{2.15}$$

O valor que maximiza $l(\beta, y)$ é também o que maximiza a função obtida pelo seu logarítmo, a chamada função de log-verosimilhança, $L(\beta, y)$, uma vez que a função logarítmo é uma função monótona. A função de log-verosimilhança é usualmente, mais adequada para cálcular o máximo, dado que permite escrever a verosimilhança como uma soma de parcelas ao invés de uma multiplicação de parcelas e, consequentemente, simplificar cálculo da derivada.

$$L(\beta, y) \quad = \quad \log\left(\prod_{i=1}^{n} f(y|x_i)\right) = \sum_{i=1}^{n} \log(f(y|x_i)) \tag{2.16}$$

Após a obtenção da função log-verosimilhança, pretende-se então calcular os seus máximos. O máximo da função é o zero da primeira derivada, sem haver necessidade de avaliar o sinal da segunda derivada, uma vez que se prova que esta tem sempre sinal negativo (Casella and Berger, 1990). Neste sentido, calculam-se as $p+1$ equações de verosimilhança correspondentes aos $p+1$ coeficientes do modelo ($p$ coeficientes associados às $p$ variáveis explicativas mais o parâmetro constante), que correspondem a derivar a equação (2.16) em ordem a cada $\beta_j$ e igualá-las a zero.

Relembrando, a *componente aleatória* de um modelo com resposta binária é:

$$Y|(X = x_i) \sim Bin(1, \pi_i) \tag{2.17}$$

e, por definição, a função de probabilidade condicionada é:

$$f(y|x_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} \tag{2.18}$$

Neste caso, a contribuição para a função de probabilidade é $\pi(x_i)$ quando $y_i = 1$ e $1 - \pi(x_i)$ caso contrário e portanto, de acordo com a equação (2.16), a função log-verosimilhança pode ser escrita como:

$$L(\beta, y) = \sum_{i=1}^{n} \{y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i))\} \tag{2.19}$$

As equações de verosimilhança assumem a seguinte forma:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \quad \text{para} \quad j = 0 \tag{2.20}$$

e

$$\sum_{i=1}^{n} x_{ij} \, [y_i - \pi(x_i)] = 0 \quad j = 1, ..., p \tag{2.21}$$

Em geral, não é possível resolver-se as equações (2.20) e (2.21) de forma explícita. Nesses casos, para obtermos as estimativas de $\beta$, $\widehat{\beta}$, é necessário recorrer-se a métodos numéricos. O mais reportado na literatura é o método iterativo dos mínimos quadrados ponderados, podendo ser consultado com detalhe em Dobson (2002), Fahrmeir and Tutz (2001), McCullagh and Nelder (1989).

### 2.2.2 Avaliação da qualidade do modelo

Depois de se obter o modelo estimado, é essencial avaliar a sua qualidade no que diz respeito, em particular, a:

- Teste ao bom ajustamento do modelo aos dados;
- Proporção de variância explicada pelo modelo;
- Significância estatística das variáveis do modelo.

Antes de analisarmos um determinado modelo, $m$, com função de verosimilhança e log-verosimilhança dadas por $l_m$ e $L_m$, respetivamente, é necessário termos em conta a existência de outros dois modelos recorrentemente utilizados em comparações (Lindsey, 1997):

- **modelo nulo:** é constituído apenas pelo termo constante $\beta_0$, não contendo nenhuma variável explicativa, apresentando a menor função de verosimilhança, $l_0$. A função de log-verosimilhança é definida por $L_0$.
- **modelo saturado:** representa exatamente a amostra, uma vez que é estimado um parâmetro para cada observação. Este modelo é o que tem uma maior função de verosimilhança, $l_s$. A sua função de log-verosimilhança assume-se como $L_s$.

**Teste ao bom ajustamento do modelo aos dados**
Um dos testes de ajuste mais usuais em MLG é a denominada *Desviância*, permitindo esta medida comparar a discrepância entre os valores observados e os valores previstos. No caso da regressão logística, o teste mais frequente para este mesmo propósito é o designado *teste de Hosmer and Lemeshow* (H&L) .

A *Desviância*, utiliza o modelo saturado, $s$, para avaliar a qualidade de ajustamento de um modelo, $m$, com base nas funções de verosimilhança:

$$D = -2ln\left(\frac{l_m}{l_s}\right) = -2(L_m - L_s) \tag{2.22}$$

onde $l_m$ e $l_s$ representam as funções de verosimilhança do modelo em estudo e do modelo saturado respetivamente e $L_m$ e $L_s$ as funções de log-verosimilhança do modelo em estudo e do modelo saturado. De acordo com as equações (2.19) e (2.22), temos então que:

$$D = -2\sum_{i=1}^{n}\left[y_i \log\left(\frac{\widehat{\pi}(x_i)}{y_i}\right) + (1-y_i)\log\left(\frac{1-\widehat{\pi}(x_i)}{1-y_i}\right)\right] \tag{2.23}$$

Esta dedução é obtida considerando por definição que no modelo saturado $\widehat{\pi}(x_i) = y_i$. Para além disto, a variável resposta assume apenas valores 0 ou 1, pelo que $l_s$=1 (Hosmer and Lemeshow, 2000). Neste seguimento, a equação (2.22) pode ser reduzida a:

$$D = -2L_m \tag{2.24}$$

De acordo com esta medida de discrepância é intuitivo concluir-se que esta se assume sempre superior ou igual a zero. Um modelo é tanto melhor quanto menor for o valor de $D$, sendo que $D = 0$, traduz um modelo com um ajustamento perfeito aos dados, como é o caso do modelo saturado.

Sob a hipótese nula:
$H_0$: o ajustamento do modelo em causa ($m$) é igual ao ajustamento do modelo saturado ($s$)
$H_1 :\sim H_0$

e tendo em conta a estatística de teste $D \sim \chi^2_{1-\alpha}(J-(p+1))$, onde $J$ representa o número de padrões de covariáveis distintas, $p$ o número de parâmetros do modelo $m$ e $\chi^2$ a estatística de qui-quadrado de Pearson usual. A hipótese nula é rejeitada com um nível de significância $\alpha$ se $D > \chi^2_{1-\alpha}(J-(p+1))$. Note-se que a *Desviância* faz o mesmo papel que a soma dos quadrados dos resíduos para a regressão linear (McCullagh and Nelder, 1989).

Este teste é reconhecido na literatura como *teste da razão de verosimilhança*, pelo facto de se considerar o quociente entre $l_m$ e $l_p$ (Agresti, 1996; Hosmer and Lemeshow, 2000; McCullagh and Nelder, 1989).

Outra medida bastante usual que nos permite compreender a discrepância entre valores observados e valores ajustados é a $\chi^2$ *de Pearson* que para a regressão logística se assume como sendo a estatística de qui-quadrado original. No entanto, estes testes não são adequados quando $J \approx n$, uma vez que apresentam problemas de convergência.

Um teste que pode ser utilizado alternativamente para contornar a situação anterior é o denominado *teste de Hosmer and Lemeshow (H&L)*, sendo o teste mais usual em regressão logística.
Sob a hipótese nula:

$H_0$ : o modelo faz um bom ajustamento aos dados
$H_1 :\sim H_0$

torna-se equivalente a reformular as hipóteses da seguinte forma:

$H_0$ : não há diferença entre os valores observados e os valores previstos
$H_1 :\sim H_0$

correspondendo assim a um teste do $\chi^2$ usual. Nesta situação, tem-se que a estatística de teste corresponde a $\widehat{C} \sim \chi^2(g-2)$, onde $g$ representa o número de grupos. Esta hipótese é rejeitada para um nível de significância $\alpha$ caso $\widehat{C} > \chi^2_{1-\alpha}(g-2)$. Destaque-se que sendo a estatística $\widehat{C}$ correspondente à estatística do qui-quadrado usual, tem limitações semelhantes. Segundo Hosmer and Lemeshow (2000), as frequências esperadas em cada grupo deverão ser pelo menos 5.
A obtenção dos $g$ grupos considerados, pressupõe algumas etapas a descrever de seguida:

**(i)** as $n$ probabilidades de sucesso para cada indivíduo ($P(Y_i = 1)$) são ordenadas crescentemente

**(ii)** as probabilidades são agora agrupadas em $g$ grupos com o mesmo número de observações ($g/n$), divididos por decis. Hosmer and Lemeshow (2000) aconselham 10 grupos, sendo que cada grupo representa um decil: por exemplo, o primeiro grupo apresenta as probabilidades mais baixas, $0-0.1$, o segundo probabilidades entre $0.1-0.2$ e assim sucessivamente.

**(iii)** obtém-se uma tabela de contingência $g \times 2$ que consta do cruzamento entre a resposta binária e os $g$ grupos. Esta estatística, $\widehat{C}$ corresponde à estatística usual do $\chi^2$ de Pearson que permite comparar a frequência dos valores observados com os valores esperados.

**Proporção de variância explicada pelo modelo**
A regressão logística não tem o equivalente direto ao coeficiente de determinação $R^2$ na regressão

linear clássica que representa a proporção de variância da variável resposta, explicada pelos preditores. Um valor de $R^2$ próximo de 100% indica que o modelo é capaz de explicar muito a variabilidade dos dados. Têm sido propostas algumas medidas, denominadas como $pseudo - R^2$ que visam ser uma medida correspondente de $R^2$ para a regressão logística (Hu et al., 2006). Estes $pseudo - R^2$ baseiam-se na comparação da função de log-verosimilhança do modelo nulo com a do modelo em estudo (Hu et al., 2006; Shtatland et al., 2002). Uma medida muito usual é o $Cox\&Snell\ R^2$:

$$Cox\&Snell\ R^2 = 1 - exp\left(\frac{-2(L_m - L_0)}{n}\right) \tag{2.25}$$

onde $L_m$ e $L_0$ são as funções de log-verosimilhança do modelo em estudo e do modelo nulo, respetivamente e $n$ é o tamanho da amostra. Este coeficiente apresenta a limitação de nunca atingir o valor 1, o que dificulta a sua interpretação. Assim, surge o $Nagelkerke's\ R^2$ como uma normalização do $Cox\&Snell\ R^2$ com a vantagem de variar entre 0 e 1.

$$Nagelkerke's\ R^2 = \frac{Cox\&Snell\ R^2}{1 - exp\left(\frac{2L_0}{n}\right)} \tag{2.26}$$

**Significância estatística das variáveis do modelo**
O *teste de Wald*, permite avaliar individualmente a significância estatística de cada coeficiente do modelo, de acordo com a hipótese

$$H_0 : \beta_j = 0,\ j = 0, ..., p \quad || \quad H_1 : \beta_j \neq 0 \tag{2.27}$$

A estatística do teste é calculada pelo quociente:

$$W = \frac{\widehat{\beta_j}}{SE(\widehat{\beta_j})} \sim N(0, 1) \tag{2.28}$$

onde $\widehat{\beta_j}$ é a estimativa de máxima verosimilhança de $\beta_j$ e $SE(\widehat{\beta_j})$ representa o erro padrão de $\widehat{\beta_j}$. A hipótese $H_0$ é rejeitada em favor de $H_1$ se $W > z_{1-\alpha/2}$, onde $z_{1-\alpha/2}$ representa o quantil $(1 - \alpha/2)$ da distribuição N(0,1) ou equivalentemente, $p - value \leq \alpha$, onde $p - value$ representa a probabilidade de se observar um valor mais extremo do que o estimado na amostra.

É importante realçar que um teste de hipóteses apenas permite refutar (ou não) a hipótese $H_0$. Neste sentido, uma forma mais informativa de se inferir sobre um parâmetro populacional é com base na informação dos **intervalos de confiança** (IC). O intervalo de $100(1 - \alpha)\%$ de confiança para um $\beta_j$, $j = 1, ..., p$ é dado por:

$$\widehat{\beta_j} \pm z_{1-\alpha/2} SE(\widehat{\beta_j}) \tag{2.29}$$

onde $z_{1-\alpha/2}$ representa o $(1 - \alpha/2)$ quantil da distribuição N(0,1). Com $100(1 - \alpha)\%$ de confiança, este intervalo conterá o valor do parâmetro $\beta_j$. Neste caso, um IC para $\beta_j$ que não contenha o valor 0 indica que $\beta_j \neq 0$ com $100(1 - \alpha)\%$ de confiança e portanto a variável a ele associada é significativa no modelo.

### 2.2.3 Interpretação do modelo

Pretende-se agora apresentar a interpretação do modelo e, em particular dos seus coeficientes. No modelo de regressão logística

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} \,, \tag{2.30}$$

cada parâmetro $\beta_j$ é interpretado como uma mudança provocada em $logit(\pi_i)$ devido ao aumento de uma unidade na j-ésima variável explicativa, $X_j$, quando as restantes variáveis se mantêm constantes (Hosmer and Lemeshow, 2000; Faraway, 2006). Para as variáveis explicativas categóricas, "uma unidade"refere-se à categoria em estudo quando comparada com a de referência. Assim, podem ocorrer três situações:

- Se $\beta_j > 0$, a probabilidade de "sucesso"aumenta com o aumento da variável explicativa $X_j$, na forma da função da figura 2.1;
- Se $\beta_j = 0$, a probabilidade de "sucesso"não depende da variável explicativa $X_j$;
- Se $\beta_j < 0$, a probabilidade de "sucesso"aumenta com a diminuição da variável explicativa $X_j$.

Para além dos coeficientes de regressão, também é habitual extrair informação dos *odds* e dos *odds ratio*. O **odds** correspondem ao quociente entre a probabilidade de "sucesso"e a probabilidade de "insucesso", isto é:

$$odds = \frac{P(Y=1)}{1-P(Y=1)} = \frac{P(Y=1)}{P(Y=0)} = \frac{\pi}{1-\pi} \tag{2.31}$$

O **odds ratio** ($OR$) é o quociente entre dois odds, permitindo comparar as probabilidades de sucesso/insucesso para dois grupos $(G_1, G_2)$ distintos:

$$OR = \frac{odds(G_1)}{odds(G_2)} = \frac{\dfrac{P(Y=1|G_1)}{1-P(Y=1|G_1)}}{\dfrac{P(Y=1|G_2)}{1-P(Y=1|G_2)}} \tag{2.32}$$

É importante frisar que os grupos em comparação $(G_1, G_2)$, são muitas vezes complementares, nomeadamente quando se comparam grupos de indivíduos com e sem uma determinada doença. No entanto, não é limitativo, pois permitem que a comparação seja sempre efetuada com uma classe definida como referência. Recorrendo aos odds ratio conseguimos obter o quão mais provável é acontecer o evento num determinado grupo em comparação com o outro.
É possível ter os seguintes casos:

- se $OR > 1$, o "sucesso"é mais provável no grupo $G_1$;
- se $OR = 1$, o "sucesso"é igualmente provável nos dois grupos;
- se $OR < 1$, o "sucesso"é mais provável no grupo $G_2$.

Os $OR$ são obtidos através dos parâmetros do modelo estimado na amostra. Esta estimação depende portanto do tipo de variável associado a cada parâmetro.

**Variável explicativa dicotómica**
Consideremos o vetor de variáveis explicativas $X = (X_1, ..., X_v, ..., X_p)$ e $X_v$ como uma variável dicotómica codificada por duas categorias, a e b. Temos então que:

$$\begin{aligned}
\log[\widehat{OR}] &= \log[\widehat{OR}(X_v = a, X_v = b)] \\
&= \log\left[\frac{\frac{\widehat{\pi}(X_v = a)}{1 - \widehat{\pi}(X_v = a)}}{\frac{\widehat{\pi}(X_v = b)}{1 - \widehat{\pi}(X_v = b)}}\right] \\
&= logit[X_v = a] - logit[X_v = b] \\
&= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + ... + \widehat{\beta}_{v-1} X_{v-1} + \widehat{\beta}_v \times a + \widehat{\beta}_{v+1} X_{v+1} + ... + \widehat{\beta}_p X_p \\
&- \widehat{\beta}_0 - \widehat{\beta}_1 X_1 - ... - \widehat{\beta}_{v-1} X_{v-1} - \widehat{\beta}_v \times b - \widehat{\beta}_{v+1} X_{v+1} - ... - \widehat{\beta}_p X_p \\
&= \widehat{\beta}_v \times (a - b) \qquad (2.33)
\end{aligned}$$

Conclui-se então que $\widehat{OR} = e^{\widehat{\beta}_v \times (a-b)}$. Note-se que a maioria das variáveis dicotómicas são codificados em 1 e 0, nomeadamente em softwares estatísticos. Assim, para este caso ($a = 1$ e $b = 0$) temos que $\widehat{OR} = e^{\widehat{\beta}_v}$. Em variáveis de natureza dicotómica, e assumindo agora que a variável apresenta codificação 1 e 0, os $\widehat{OR}$ são interpretados como: é mais/menos provável $e^{\widehat{\beta}_v}$ vezes a ocorrência do evento para os objetos com $X_v = 1$ em comparação com os que assumem $X_v = 0$, quando as restantes variáveis se mantêm constantes.

**Variável explicativa politómica**
Quando a variável $X_v$ assume ser politómica com $k$ categorias, esta é transformada em $k-1$ dummys e o processo é análogo ao da variável dicotómica. Esta codificação permite a comparação de categorias duas a duas, onde cada categoria é sempre comparada com uma categoria de referência.

**Variável explicativa contínua**
Assumindo-se como contínua a variável $X_v$, a interpretação do respetivo $\widehat{OR}$ varia de acordo com as unidades da variável. Usualmente, as variáveis contínuas são interpretadas de acordo com o aumento de uma unidade, por exemplo o peso: "à medida que aumenta uma grama de peso"ou a idade gestacional: "à medida que aumenta uma semana de gestação". Contudo, esta abordagem de uma unidade nem sempre é a mais correta, e, como tal, consideremos que a variável $X_v$ aumenta de $c$ unidades, ou seja, $X = (X_1, ..., X_v + c, ..., X_p)$. Temos então que:

$$\begin{aligned}
\log[\widehat{OR}(X_v + c, X_v)] &= \log\left[\frac{odds(X_v + c)}{odds(X_v)}\right] \qquad (2.34) \\
&= logit[X_v + c] - logit[X_v] = (X_v + c)\beta_v - \beta_v X_v \\
&= c\beta_v \qquad (2.35)
\end{aligned}$$

Conclui-se que $\widehat{OR} = e^{c\beta_v}$, traduzindo o $OR$ para o "sucesso"por aumento de $c$ unidades na variável $X_v$ e considerando as restantes covariáveis constantes.

Analogamente ao que se efetua para os $\beta_j$, a significância das variáveis no modelo pode também ser obtida com recurso aos $\widehat{OR}$. O teste descrito pela equação (2.27) para $\beta_j$ é equivalente a testar a hipótese seguinte :

$$H_0 : OR = 1 \quad || \quad H_1 : OR \neq 1 \qquad (2.36)$$

uma vez que se provou que $OR = e^{\beta_j}$. Está-se a testar a hipótese de a probabilidade de sucesso do evento ser igualmente provável em ambos os grupos considerados (OR=1). Isto significa que, pretendemos rejeitar a hipótese nula para que uma variável seja considerada significativa.
Vimos na secção 2.2.2 que através dos coeficientes estimados na amostra, é-nos possível inferir sobre a população e determinar quais as variáveis significativas nesta. Este procedimento obtém-se através da aplicação do *teste de Wald* e de intervalos de confiança (IC) para os respetivos coeficientes de regressão, $\beta_j$ (equações (2.27) e (2.29)).
Um intervalo com $100(1-\alpha)\%$ de confiança para os OR é então dado por:

$$e^{\widehat{\beta}_j \pm z_{1-\alpha/2} SE(\widehat{\beta}_j)} \tag{2.37}$$

onde $z_{1-\alpha/2}$ representa o $(1-\alpha/2)$ quantil da distribuição $N(0,1)$ e $SE$ o erro padrão associado ao coeficiente. Note-se que este IC corresponde a exponencializar o IC referente à equação (2.29). Seguindo a analogia anterior, para uma variável explicativa $X_j$ ser considerada significativa, o intervalo de confiança do respetivo OR não poderá conter o valor 1.

*Exemplo:* Considere-se que se pretende determinar se a variável ETT Resuscitation "necessidade de intubação"é considerada um fator de risco de mortalidade para recém nascidos prematuros extremos. Esta variável apresenta-se como dicotómica, onde 1 corresponde a necessidade de intubação e 0 o seu complementar. A estimativa do OR para o evento mortalidade foi de:

$$\widehat{OR}[ETTResuscitation] = \frac{\dfrac{P(morrer|ETTResuscitation)}{P(sobreviver|ETTResuscitation)}}{\dfrac{P(morrer|\overline{ETTResuscitation})}{P(sobreviver|\overline{ETTResuscitation})}} = 2.04$$

e o intervalo de confiança a 95% para o $OR$ populacional foi de $[1.03, 4.03]$, indicando que a necessidade de intubação está associada a um aumento do risco de mortalidade, quando comparada com a não necessidade da mesma.

Assumindo que IC= $[l, u]$, existem 3 alternativas possíveis de resultados semelhantes aos da página 21:

- Se $[l, u] \subset 1$, a mortalidade é **igualmente provável** de ocorrer nos dois grupos;
- Se $[l, u] \subset [0, 1[$, a necessidade de intubação de um recém nascido prematuro está associada a uma **diminuição** do risco de mortalidade;
- Se $[l, u] \subset ]1, +\infty[$, a necessidade de intubação está associada a um **aumento** do risco de mortalidade.

## 2.3 Preparação da amostra e desenho do estudo

Nesta secção descreve-se o procedimento de preparação da amostra e desenho do estudo para a identificação de fatores de risco precoces de mortalidade e morbilidade.

Para este estudo foram excluídos os stillbirths (SB) e os delivery room death (DRD), como já mencionado anteriormente (Capítulo1, subsecção 1.3.1).

Foi selecionado um conjunto de variáveis da tabela 1.3 recolhidas praticamente no primeiro dia de vida de um recém nascido e que são consideradas como potenciais fatores de risco. Estes fatores

de risco foram divididos em dois subconjuntos − *protocolares* e *intrínsecos*− de acordo com a sua origem. Por um lado, os fatores protocolares têm o propósito de avaliar a importância das intervenções médicas perinatais (por exemplo, o uso de corticoides (Antenatal Steroids)). Por outro lado, foram considerados fatores intrínsecos que dizem respeito a caraterísticas fisiológicas de cada recém nascido (por exemplo, o género (Gender) e o peso (Weight)).
Assim, considerou-se:

- $X_j$ como o potencial fator de risco, $j = 1, ..., p$;
- $Y$ a variável resposta dicotómica que conforme se trata de análise de mortalidade ou de morbilidade, é definida por

$$\text{Mortalidade} \quad Y = \begin{cases} 1, & \textbf{Dead} \text{ (se o bebé morre)} \\ 0, & \textbf{Alive} \text{ (se o bebé sobrevive)} \end{cases}$$

$$\text{Morbilidade} \quad Y = \begin{cases} 1, & \textbf{Severe} \text{ (se o bebé possui anomalias severas)} \\ 0, & \textbf{Non Severe} \text{ (se o bebé não possui anomalias severas)} \end{cases}$$

O valor $Y = 1$ modelo o outcome mais grave, **Dead** ou **Severe**, consoante o estudo é de mortalidade ou morbilidade.
Os fatores de risco foram identificados através do método de regressão logística binário apresentado na secção 2.2, onde se estimaram odds ratio para cada situação considerada, bem como o respetivo intervalo de confiança. Para cada um dos dois estudos (mortalidade e morbilidade) foram considerados três modelos, como resume a tabela 2.1.

Tabela 2.1: Modelos considerados para a identificação de fatores de risco precoces, onde $\pi$ representa a probabilidade de sucesso do evento (mortalidade ou morbilidade) e $\beta_j$ os coeficientes de regressão associados às respetivas variáveis $X_j$.

| Denomination | Model |
|---|---|
| Crude | $logit(\pi) = \beta_0 + \beta_j X_j, \ \ j = 1, ..., p$ |
| Adjusted | $logit(\pi) = \beta_0 + \beta_j X_j + \beta_{GA} X_{GA}, X_j \neq X_{GA}$ |
| Multivariate | $logit(\pi) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \beta_{GA} X_{GA}, X_j \neq X_{GA}$ |

Inicialmente, os odds ratio foram estimados para cada fator de risco através da regressão logística simples (modelo Crude, Tabela 2.1). Posteriormente, os odds ratio foram ajustados para a idade gestacional (GA), com o intuito de se descartarem possíveis associações entre os fatores de risco e o protocolo clínico ou intervenções médicas, dado que ambos, na maioria das vezes, são procedidos com recurso a esta variável (modelo Adjusted, Tabela 2.1). Por fim, estudou-se o efeito conjunto de diversos fatores de risco recorrendo à regressão logística multivariada, incluindo a GA como covariável (modelo Multivariate, Tabela 2.1).
Note-se que a GA é também considerada um fator de risco em toda a análise dos fatores intrínsecos e quando avaliado o modelo multivariado protocolar.

O modelo multivariado foi obtido considerando dois procedimentos do tipo *stepwise*, baseados no teste de Wald para um nível de significância de 5%:

- **Forward:** inicia-se com o modelo nulo, constituído apenas pelo termo constante. As variáveis vão sendo adicionadas sequencialmente até ser possível melhorar o desempenho do modelo.

Este tipo de seleção termina quando uma variável adicionada ao modelo não é significativa de acordo com o nível de significância pré-definido, não sendo introduzidas mais nenhuma variável ao modelo.

- **Backward:** o processo é contrário ao anterior. Inicia-se com o modelo completo, onde estão todas as variáveis candidatas a preditoras. Posteriormente, estas variáveis são eliminadas sucessivamente do modelo, até que todas as variáveis presentes sejam significativas, de acordo com 0.05, neste caso.

Para cada um dos procedimentos (Forward e Backward) obtém-se um modelo. A utilização de dois procedimentos diferentes teve como propósito adquirir mais confiança relativamente ao modelo multivariado. O bom ajustamento dos modelos multivariados foi avaliado pelo teste de H&L. No que diz respeito à qualidade destes modelos, a avaliação foi efetuada através das estatísticas Cox&Snell $R^2$e Nagelkerke $R^2$, conhecidas como sendo uma aproximação da proporção de variância da variável resposta, explicada pelos respetivos modelos.

## 2.4 Resultados e Discussão

Esta secção de resultados e discussão encontra-se dividida em duas partes, uma dedicada ao estudo de mortalidade (subsecção 2.4.1) e outra dedicada ao estudo de morbilidade (subsecção 2.4.2).

### 2.4.1 Fatores de risco de mortalidade

A figura 2.2 apresenta os resultados da análise individual de cada fator de risco protocolar e intrínseco associado à mortalidade ao um ano de idade. Nesta etapa, trabalhou-se no sentido da visualização, sendo que o desenvolvimento integral da figura 2.2 teve como objetivo a interpretabilidade imediata dos resultados. No eixo horizontal estão representados os valores possíveis para os $\widehat{OR}$ e respetivos intervalos de confiança, sendo que o eixo vertical representa as variáveis explicativas (fatores de risco) do problema. O esquema de cores utilizado foi escolhido de modo a que seja possível depreender a gravidade do fator de risco relativamente à mortalidade (neste caso). A ideia base foi criar uma figura em que as cores atribuídas fossem semelhantes às de um semáforo (à exceção da cor cinzenta em substituição da amarela). Assim, todos os fatores de risco não significativos para a mortalidade, estão coloridos a cinzento, dando a ideia de "desprezo"; aqueles que contribuem para uma maior sobrevivência, apresentam-se a verde, representando um resultado favorável e otimista; e, por fim, os vermelhos ilustram os fatores de risco associados a um resultado negativo, ou seja, a um aumento de mortalidade.

Referente aos fatores protocolares (Figura 2.2), a regressão logística simples indica que a época (Epoch), a gravidez vigiada (Pregnancy Surveillance), o parto iatrogénico (Iatrogenic Delivery) e o surfactante (Surfactant) não se traduzem em fatores de risco significativos para a mortalidade. Realce-se que o fator de risco época foi introduzido nesta investigação, na medida em que alguns destes anos traduzem a introdução ou mudança de intervenções médicas, de que a MJD foi alvo. Porém, essas oscilações não se relacionam significativamente com a mortalidade.

Por outro lado, o nascimento num hospital especializado (Inborn at MJD), o uso de corticoides (Antenatal Steroids) e o parto por cesariana (Caesarean Delivery) estão associados a um baixo risco de mortalidade, em conformidade com um estudo sueco em condições análogas (The Express Group Members, 2010).

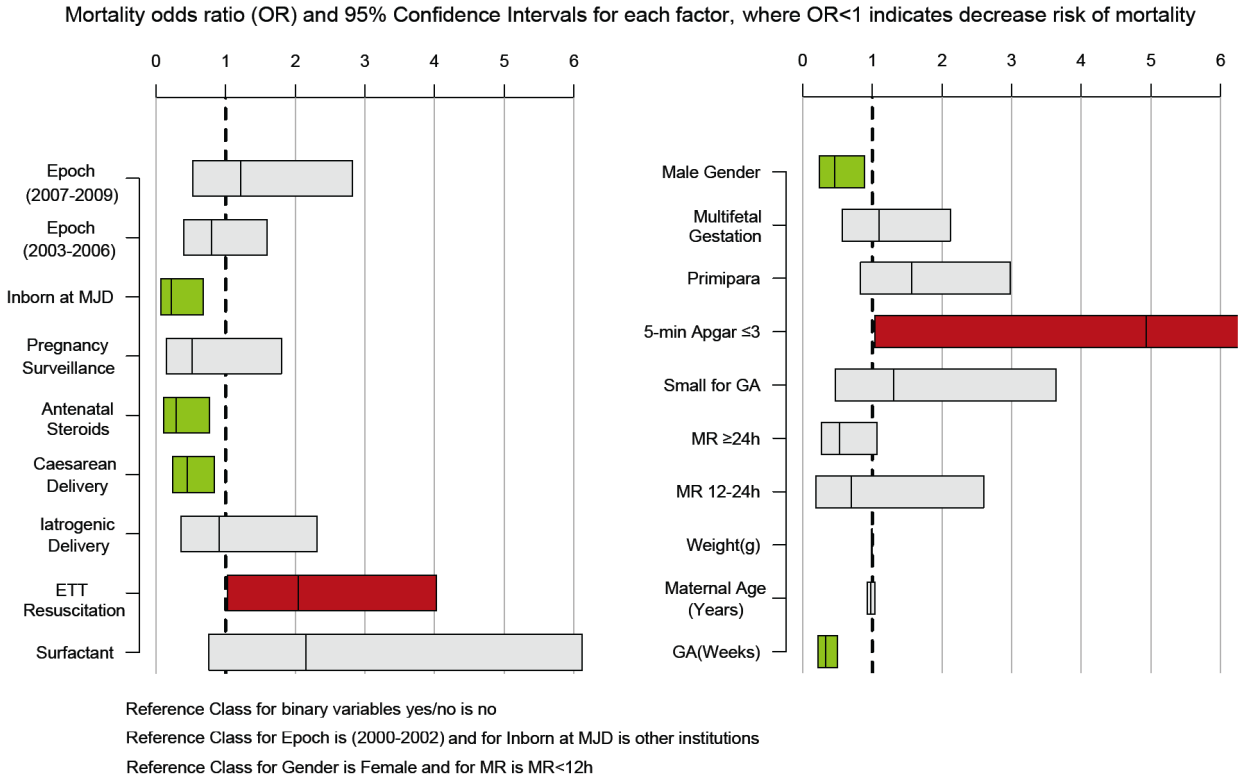



Figura 2.2: *OR* para a mortalidade e respetivos intervalos de confiança a 95% para cada fator de risco protocolar (esquerda) e intrínseco (direita) utilizando regressão logística univariada (modelo Crude, Tabela 2.1).

O facto de o local de nascimento demonstrar ser um fator significativo, indica que os prematuros nascidos na MJD poderão beneficiar de cuidados essenciais mais atempadamente em relação aos recém-nascidos que necessitam de serem transferidos de outras instituições. Para além disto, é notório que o transporte agrava o prognóstico de um recém nascido prematuro, na medida em que estes são demasiado imaturos e portanto, muito suscetíveis a diversos problemas envolventes (Araújo et al., 2011). Logo, o nascimento num hospital considerado especializado na área de neonatologia é uma mais valia em recém nascidos prematuros extremos (Arad et al., 2008; The Express Group Members, 2010).

Relativamente aos corticoides, este resultado poderá ser explicado na medida em que este fármaco é éticamente recomendável de ser administrado em gravidezes que apresentem risco de parto prematuro, nomeadamente entre as 24 e as 34 semanas de gestação, com o intuito de desenvolver a maturidade pulmonar da criança. No entanto, investigações atuais garantem que a terapia de corticoides em mães de crianças com idades gestacionais inferiores a 25 semanas está associada a uma diminuição do risco de mortalidade quando comparados a recém nascidos que não usufruíram do fármaco (Carlo et al., 2011; Smith et al., 2012). Assim, é destacado o facto de esta intervenção protocolar permitir um melhor desenvolvimento do aparelho respiratório e consequentemente um aumento de sobrevivência.

Quanto ao parto por cesariana nesta idade gestacional (<27 semanas de gestação), está comprovado na literatura que pode diminuir o risco de mortalidade em diversas situações, nomeadamente em GA<26 semanas (grande parte dos recém nascidos prematuros da base de dados), gestações múltiplas e apresentação de pelve (quandos os fetos estão sentados) (Berger et al., 2011). Na

publicação de Wang et al. (2011), estes autores evidenciam ainda que alguns profissionais de saúde defendem que a cesariana deve ser utilizada livremente devido à fragilidade destas crianças e em vista a minimizar possíveis traumatismos.

Na população de recém nascidos prematuros extremos da MJD, a necessidade de intubação (ETT Resuscitation), julgada pela opinião médica, revela ser um fator de risco que contribui para o aumento da mortalidade. Este resultado está de acordo com o esperado, dado que ETT Resuscitation é um experimento não aleatório, isto é, trata-se de uma intervenção médica utilizada apenas em recém nascidos prematuros que se encontram em condições bastante críticas e, portanto, com pouca probabilidade de sobrevida.

A tabela 2.2 apresenta as estimativas dos $OR$ e os respetivos intervalos de confiança para cada abordagem, onde é possível observar-se que depois de ajustado à GA, de todos os fatores de risco considerados apenas Inborn at MJD continua relacionado com uma diminuição da mortalidade, indiciando que a GA não se associa ao nascimento nesta unidade hospitalar.

Finalmente, numa análise multivariada das intervenções protocolares (Tabela 2.2c), verificou-se que a não necessidade de ETT Resuscitation, Inborn at MJD e o aumento da GA são fatores de risco que decrescem significativamente o risco de mortalidade. A inclusão da variável GA como significativa no modelo de mortalidade, já era esperado, visto que esta variável é considerada a mais relevante neste tipo de estudos (Boussicault et al., 2012; The Express Group Members, 2010).

Em relação aos fatores intrínsecos, a figura 2.2 indica que o risco de mortalidade diminui para o sexo masculino (Male Gender), 5-min Apgar>3 e também com o aumento do Weigth e da GA.

A tabela 2.3 exibe os valores dos $OR$ estimados e respetivos intervalos de confiança para cada abordagem baseada em fatores intrínsecos, onde se observa que quando ajustado à GA, 5-min Apgar$\geq$3 revelou não ser um fator significativo. Já o aumento do peso, o género masculino e a MR$\geq$24 h quando comparada com MR$<$ 12 h, indicam ser fatores de risco significativos de baixa mortalidade (Tabela 2.3b).

Por fim, num efeito conjunto dos fatores de risco (Tabela 2.3c), resultou que o aumento do Weight e da GA refletem ser fatores de risco intrínsecos significativos para a diminuição do risco de mortalidade, em concordância com inúmera literatura de estudos semelhantes (Boussicault et al., 2012; The Express Group Members, 2010). É no entanto intuitivo compreender que o aumento do peso e da idade gestacional está relacionado com uma maior sobrevivência por parte de um recém nascido prematuro extremo, uma vez que este se encontra mais desenvolvido e com mais maturidade estando, portanto, menos sujeito a complicações.

**Índice de apgar**

Nesta investigação, um apgar inferior ou igual a 3 aos 5 minutos indica que o recém nascido prematuro está numa situação muito crítica, aumentando significativamente o risco de morte. Esta referência do apgar e respetiva associação com a mortalidade corrobora com o trabalho de (Wang et al., 2011). Já quando ajustado à GA, este deixa de ser um fator de risco significativo, possivelmente pelo facto de ser conhecido que com o avanço da GA o índice de apgar aumenta, mostrando uma associação evidente entre estas duas variáveis (The Express Group Members, 2010). Todavia, tal evidência não foi comprovada nesta investigação, dado que não foram detetadas

diferenças significativas para a GA[1] ($p - value$=0.056).

Tabela 2.2: OR para a mortalidade e respetivo intervalo de confiança a 95% para os fatores de risco protocolares.

| | | | | **Odds Ratio (95 % of Confidence Interval)** | | |
|---|---|---|---|---|---|---|
| | | **All n=167*** | **Dead n=87** | **(a) Crude** | **(b) Adjusted** | **(c) Multivariate** |
| Epoch | (2007-2009) | 36 | 21 | **1.22**(0.53,2.82) | **1.17**(0.45,3) | |
| | (2003-2006) | 73 | 35 | **0.8**(0.4,1.6) | **0.85**(0.4,1.81) | – – – – – – – |
| | (2000-2002) | 58 | 31 | 1[Reference] | 1[Reference] | |
| Inborn delivery at | MJD | 146 | 70 | **0.22**(0.07,0.68) | **0.23**(0.07,0.77) | **0.21**(0.06,0.71) |
| | Others | 21 | 17 | 1[Reference] | 1[Reference] | 1[Reference] |
| Pregnancy Surveillance | Yes | 155 | 79 | **0.52**(0.15,1.8) | **1.12**(0.27,4.6) | – – – – – – – |
| | No | 12 | 8 | 1[Reference] | 1[Reference] | |
| Antenatal Steroids | Yes | 142 | 68 | **0.29**(0.11,0.77) | **0.42**(0.14,1.24) | – – – – – – – |
| | No | 25 | 19 | 1[Reference] | 1[Reference] | |
| Caesarean Delivery | Yes | 96 | 42 | **0.45** (0.24,0.84) | **0.87** (0.42,1.81) | – – – – – – – |
| | No | 41 | 45 | 1[Reference] | 1[Reference] | |
| Iatrogenic Delivery | Yes | 20 | 10 | **0.91**(0.36,2.31) | **1.13**(0.4,3.16) | – – – – – – – |
| | No | 147 | 77 | 1[Reference] | 1[Reference] | |
| ETT Resuscitation | Yes | 119 | 68 | **2.04**(1.03,4.03) | **2.04**(0.96,4.33) | **2.27**(1.04,4.99) |
| | No | 48 | 19 | 1[Reference] | 1[Reference] | 1[Reference] |
| Surfactant | Yes | 150 | 81 | **2.15**(0.76,6.12) | **2.34**(0.75,7.28) | – – – – – – – |
| | No | 17 | 6 | 1[Reference] | 1[Reference] | |
| GA (Weeks) | | – – – | – – – | – – – – – – – | – – – – – – – | **0.37**(0.25,0.55) |

Binary Logistic Regression: (a) simple;(b) adjusted for GA;(c) multivariate with stepwise procedure including GA;
*=205 cases excluding stillbirths, delivery room death and missing values.

Tabela 2.3: OR para a mortalidade e respetivo intervalo de confiança a 95% para os fatores de risco intrínsecos.

| | | | | **Odds Ratio (95 % of Confidence Interval)** | | |
|---|---|---|---|---|---|---|
| | | **All n=153*** | **Dead n=81** | **(a) Crude** | **(b) Adjusted** | **(c) Multivariate** |
| Gender | Male | 89 | 40 | **0.46**(0.24,0.89) | **0.43**(0.20,0.89) | – – – – – – – |
| | Female | 64 | 41 | 1[Reference] | 1[Reference] | |
| Multifetal Gestation | Yes | 57 | 31 | **1.10**(0.57,2.12) | **1.28**(0.61,2.68) | – – – – – – – |
| | No | 96 | 50 | 1[Reference] | 1[Reference] | |
| Primipara | Yes | 75 | 44 | **1.57**(0.83,2.98) | **1.37**(0.67,2.81) | – – – – – – – |
| | No | 78 | 37 | 1[Reference] | 1[Reference] | |
| 5-min Apgar $\leq 3$ | Yes | 12 | 10 | **4.93**(1.04,23.31) | **3.66**(0.68,19.61) | – – – – – – – |
| | No | 141 | 71 | 1[Reference] | 1[Reference] | |
| Small for GA | Yes | 17 | 10 | **1.31**(0.47,3.64) | **2.26**(0.73,6.95) | – – – – – – – |
| | No | 136 | 71 | 1[Reference] | 1[Reference] | |
| MR | $\geq$ 24h | 51 | 22 | **0.53**(0.27,1.07) | **0.42**(0.19,0.93) | |
| | 12-24h | 10 | 5 | **0.70**(0.19,2.60) | **0.70**(0.17,2.89) | – – – – – – – |
| | $<$ 12h | 92 | 54 | 1[Reference] | 1[Reference] | |
| Weight (g) | | – – – | – – – | **0.993**(0.991,0.996) | **0.996**(0.993,0.999) | **0.996**(0.993,0.999) |
| Maternal Age (Years) | | – – – | – – – | **0.98**(0.93,1.04) | **1.01**(0.95,1.08) | – – – – – – – |
| GA (Weeks) | | – – – | – – – | **0.33**(0.22,0.50) | – – – – – – – | **0.44**(0.28,0.70) |

Binary Logistic Regression: (a) simple;(b) adjusted for GA;(c) multivariate with stepwise procedure including GA;
*=205 cases excluding stillbirths, delivery room death and missing values.

[1]Utilizou-se o teste bilateral de Mann Whitney U para testar a igualdade de medianas de GA em amostras independentes.

**Rutura de membrana**
O resultado obtido nesta investigação relativamente a este fator de risco revela estar de acordo com o estudo de Blumenfeld et al. (2010) onde se conclui que a rutura de membrana prolongada (no artigo em questão definida como superior a 18 horas antes do parto) está associada a um aumento da sobrevivência em recém nascidos prematuros entre as 24 e as 26 GA, quando comparadas com ruturas de membranas recentes ou mesmo inexistência de rutura. Embora o artigo publicado recentemente por Blumenfeld et al. (2010) corrobore os resultados obtidos e indique que a rutura de membrana não se associa a um aumento de mortalidade neonatal, outra literatura contrapõe tal resultado, referindo que as ruturas de membrana poderão resultar em diversas complicações para a mãe e para o bébé, de acordo com o tempo em que a membrana permanece rota antes do parto (Muris et al., 2007; Waters and Mercer, 2009). Deste modo, foi efetuada uma análise deste fator de risco, com o propósito de tentar compreender o sucedido neste estudo.

- MR versus Antenatal Steroids

Uma primeira abordagem passou por verificar se os bebés que apresentavam MR≥24 h teriam tido uma administração de Antenatal Steroids superior à dos bebés com MR <12 h, uma vez que esta intervenção obstétrica é aconselhável quando há um grande risco de parto prematuro e consequentemente, no caso de ruturas de membranas prematuras (Waters and Mercer, 2009). Blumenfeld et al. (2010) especulam que a redução de mortalidade associada às ruturas de membranas prolongadas poderá ser atribuída à exposição de corticoides por mais tempo. Não obstante, neste estudo não existe evidência estatística de associação entre MR e Antenatal Steroids ($p-value$=0.47 no teste do qui-quadrado[2]). No entanto, há uma limitação subjacente a este resultado, pois a variável Antenatal Steroids apenas distingue se um bébé usufruiu ou não do tratamento (Tabela 1.3), mas esta intervenção assume diversos ciclos de administração. Uma divisão desta variável distinguindo o tratamento de ciclos completos (intervenção mais prolongada) e incompletos de Antenatal Steroids talvez fosse mais adequada para se verificar se eventualmente o menor risco de morte em recém nascidos prematuros extremos com MR≥24h, está associado a um maior número de ciclos completos de Antenatal Steroids, tratamento mais eficaz do que o incompleto.

- MR versus GA

Como já constatado, a GA é um fator de risco de mortalidade (menor GA maior risco de mortalidade) (Tabela 2.3a, Figura 2.2). Neste seguimento, uma segunda abordagem traduziu-se na verificação de possível associação entre a MR e a GA, focando-nos na seguinte questão:
*Será que os bébés com MR≥24h têm GA superiores aos que apresentam MR<12h, e por isso, maior sobrevivência?*
Esta hipótese foi descartada pois repare-se que quando ajustada à GA a MR≥ 24h é significativa, indicando que estes dois fatores de risco não são totalmente correlacionados.

- MR versus Infection

Uma outra interpretação relativamente ao resultado de que MR≥ 24h está associada a uma diminuição de mortalidade quando comparada com MR>12h, seria realizar um estudo incidente sobre uma possível associação entre a MR e a presença ou não de infeção. Esta hipótese surge, pois o maior problema de se ter membranas rotas durante muito tempo é a grande probabilidade de se contrair infeções, quer pela mãe quer pelo feto (Muris et al., 2007; Waters and Mercer, 2009). Porém, não foi possível concluir acerca desta questão, pois a variável Infection apenas indica se existiu ou não infeção, não distinguindo a origem de tal.

[2]Teste utilizado para testar independência, $H_0$: MR é independente de Antenatal Steroids || $H_1 :\sim H_0$.

**Género**

Neste estudo, o género masculino revelou ser um fator protetor quando avaliado pela regressão logística simples, contrariamente ao que é reportado na literatura há mais de duas décadas (Bhaumik et al., 2004; Peacock et al., 2012; Stevenson et al., 2000). Embora todos estes artigos indiquem uma notória vantagem de sobrevivência para o género feminino, Bhaumik et al. (2004) destacam uma diminuição da mortalidade de rapazes em relação a raparigas, nos bebés nascidos antes das 30 semanas de gestação, diminuindo assim a diferença existente entre géneros. Além disto, Bhaumik et al. (2004) referem que os rapazes são mais suscetíveis a problemas respiratórios (daí morrerem mais do que as raparigas pois, antigamente, os problemas respiratórios eram considerados a principal causa de morte dos bébés prematuros) mas simultaneamente têm uma resposta mais favorável aos tratamentos por Antenatal Steroids (tornam os pulmões mais maduros). Por este motivo, a doença respiratória deixa de ser um problema e a mortalidade dos rapazes aproxima-se da das raparigas. Recentemente, também Peacock et al. (2012) referem que a introdução destas novas intervenções, como é o caso dos corticoides, poderá ser das principais causas para que a diferença de mortalidade entre os sexos esteja menos acentuada, embora nunca se tenha concluído maior sobrevivência para o género masculino.

Neste estudo, há evidência estatística de que rapazes têm um menor risco de mortalidade (Tabela 2.3), que desapareceu no modelo multivariado. Por este motivo, questiona-se a associação com outros fatores de risco.

A figura 2.3 mostra a caixa de bigodes de GA e Weight, distinguindo raparigas e rapazes.

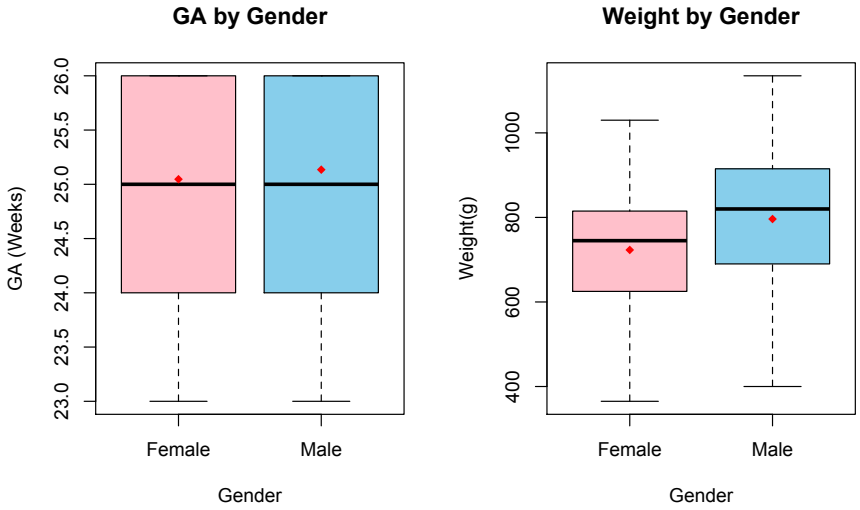



Figura 2.3: Boxplots para comparação das medianas populacionais da idade gestacional e do peso de acordo com o género. Os pontos em cada caixa representam o valor médio amostral.

A comparação das medianas amostrais indicou não existirem diferenças significativas para a GA ($p-value$=0.53) mas sim para Weight ($p-value$=0.001), isto é, rapazes da mesma idade gestacional são mais pesados do que as raparigas[3].

A bibliografia médica indica que os rapazes têm uma média de peso à nascença mais elevada do

[3]Utilizou-se o teste de Mann-Whitney U para testar a igualdade de medianas em amostras independentes; bilateral para GA e unilateral para Weight.

que as raparigas, conforme foi evidenciado nos resultados acima (Bhaumik et al., 2004; Peacock et al., 2012; Stevenson et al., 2000). Porém, os rapazes são mais propícios a nascerem prematuros, o que não se verifica na nossa população em estudo.

No entanto, o resultado obtido está de acordo com o facto de para a mesma GA, os bébés prematuros do género masculino serem mais pesados do que os do género feminino. O facto de o Gender não estar associado com a GA justifica o porquê de este ter sido identificado como fator de risco no modelo ajustado (Tabela 2.3). Finalmente, a associação do Weight com o Gender justifica ainda a questão do Gender ter saído do modelo multivariado, uma vez que o peso é mais associado com a mortalidade do que o género. A circunstância de o Weight ser considerado um fator de risco (maior peso menor mortalidade) e ter sido encontrada com evidência estatística uma associação entre Gender/Weight, e as GA serem semelhantes, é natural que neste estudo, as raparigas tenham maior risco de mortalidade.

Os modelos multivariados referentes a cada análise protocolar e intrínseca (Tabela 2.2c, Tabela 2.3c) revelaram ser os mesmos quando considerados os dois procedimentos stepwise: forward e backward. Adicionalmente, ambos mostraram fazer um ajustamento adequado dos dados quando acedidos pelo teste de H&L ($p - value$=0.52 e 0.89, respetivamente). O valor da estatística de Cox&Snell corresponde a 22.9% para os fatores protocolares e 24.1% para os fatores intrínsecos. Recorrendo à estatística de Nagalkerke, os valores resultantes foram de 30.5% (protocolares) e 32.1% (intrínsecos), o que sugere que a maioria da variância de mortalidade continua por explicar.

### 2.4.2 Fatores de risco de morbilidade

Os fatores de risco de morbilidade foram identificados seguindo a mesma metodologia usada para mortalidade. As observações presentes neste estudo reportam-se a um subconjunto de recém nascidos prematuros extremos da amostra de mortalidade, ou seja, dizem respeito aos recém nascidos prematuros extremos que sobreviveram ao fim de um ano. Nestas condições, o tamanho da amostra é bastante reduzido, existindo no máximo $n = 74$ bébés.

A tabela 2.4 apresenta os resultados para os fatores de risco protocolares, onde é possível constatar que não foram identificados fatores de risco precoces de morbilidade. Os intervalos de confiança para os odds ratio são muito longos, desfecho já esperado como consequência do reduzido tamanho da amostra.

De forma equivalente, para os fatores intrínsecos também não se identificaram fatores de risco precoces associados significativamente a um desenvolvimento neurológico severo (Tabela 2.5). Destacaram-se, no entanto, os intervalos de confiança de menor amplitude e em torno do valor 1 das variáveis Weight e Maternal age. Outro destaque surge para a variável Multifetal Gestation que, embora com intervalos de confiança muito longos, o valor 1 aproxima-se muito da banda inferior do intervalo de confiança, sugerindo que um aumento do tamanho da amostra poderia conduzir à significância estatística deste fator de risco.
Acresce da limitação da dimensão da amostra a impossibilidade da estimação correta do OR referente ao fator de risco 5-min Apgar$\leq$3, uma vez que existe, um único recém nascido com 5-min Apgar$\leq$3. Relembre-se que os bébés do estudo de morbilidade são um subconjunto de bébés que sobreviveram ao fim de um ano e 5-min Apgar$\leq$3 foi associado significativamente a um aumento de mortalidade (subsecção 2.4.1), pelo que é natural que bébés que sobreviveram ao primeiro ano de vida tenham apgar superiores.

Tabela 2.4: OR para a morbilidade e respetivo intervalo de confiança a 95% para os fatores de risco protocolares.

| | | | | Odds Ratio (95 % of Confidence Interval) | | |
|---|---|---|---|---|---|---|
| | | **All n=74*** | **Severe n=15** | **(a) Crude** | **(b) Adjusted** | **(c) Multivariate** |
| Epoch | (2007-2009) | 13 | 4 | **2.96**(0.55,16.08) | **3.05**(0.55,16.8) | |
| | (2003-2006) | 38 | 8 | **1.78**(0.42,7.52) | **1.94**(0.45,8.35) | – – – – – – – |
| | (2000-2002) | 23 | 3 | 1[Reference] | 1[Reference] | |
| Inborn delivery at | MJD | 71 | 14 | **0.49** (0.04,5.81) | **0.46**(0.04,5.7) | – – – – – – – |
| | Others | 3 | 1 | 1[Reference] | 1[Reference] | |
| Pregnancy Surveillance | Yes | 71 | 14 | **0.49**(0.04,5.81) | **0.3**(0.02,4.37) | – – – – – – – |
| | No | 3 | 1 | 1[Reference] | 1[Reference] | |
| Antenatal Steroids | Yes | 69 | 14 | **1.02** (0.11,9.84) | **0.81**(0.08,8.32) | – – – – – – – |
| | No | 5 | 1 | 1[Reference] | 1[Reference] | |
| Caesarean Delivery | Yes | 51 | 12 | **2.05** (0.52,8.11) | **1.54**(0.32,7.49) | – – – – – – – |
| | No | 23 | 3 | 1[Reference] | 1[Reference] | |
| Iatrogenic Delivery | Yes | 10 | 1 | **0.4**(0.05,3.4) | **0.37**(0.04,3.2) | – – – – – – – |
| | No | 64 | 14 | 1[Reference] | 1[Reference] | |
| ETT Resuscitation | Yes | 45 | 11 | **2.02**(0.58,7.1) | **1.91**(0.54,6.78) | – – – – – – – |
| | No | 29 | 4 | 1[Reference] | 1[Reference] | |
| Surfactant | Yes | 63 | 14 | **2.86**(0.34,24.27) | **2.98**(0.35,25.53) | – – – – – – – |
| | No | 11 | 1 | 1[Reference] | 1[Reference] | |
| GA (Weeks) | | – – – | – – – | – – – – – – – | – – – – – – – | – – – – – – – |

Binary Logistic Regression: (a) simple;(b) adjusted for GA;(c) multivariate with stepwise procedure including GA; *=74 (all possible observations into the analysis)

Tabela 2.5: OR para a morbilidade e respetivo intervalo de confiança a 95% para os fatores de risco intrínsecos.

| | | | | Odds Ratio (95 % of Confidence Interval) | | |
|---|---|---|---|---|---|---|
| | | **All n=67*** | **Severe n=13** | **(a) Crude** | **(b) Adjusted** | **(c) Multivariate** |
| Gender | Male | 46 | 10 | **1.67**(0.41,6.82) | **1.66**(0.4,6.87) | – – – – – – – |
| | Female | 21 | 3 | 1[Reference] | 1[Reference] | |
| Multifetal Gestation | Yes | 25 | 8 | **3.48**(0.99,12.23) | **2.94**(0.76,11.26) | **3.48**(0.991,12.23) |
| | No | 42 | 5 | 1[Reference] | 1[Reference] | 1[Reference] |
| Primipara | Yes | 29 | 7 | **1.7**(0.5,5.74) | **1.83**(0.53,6.3) | – – – – – – – |
| | No | 38 | 6 | 1[Reference] | 1[Reference] | |
| 5-min Apgar$\leq$ 3 | Yes | 1 | 0 | **0**(0,Inf) | **0**(0,Inf) | – – – – – – – |
| | No | 66 | 13 | 1[Reference] | 1[Reference] | |
| Small for GA | Yes | 7 | 1 | **0.67**(0.07,6.07) | **0.5**(0.05,4.65) | – – – – – – – |
| | No | 60 | 12 | 1[Reference] | 1[Reference] | |
| MR | $\geq$ 24h | 25 | 2 | **0.23**(0.05,1.18) | **0.27**(0.05,1.36) | – – – – – – – |
| | 12-24h | 5 | 1 | **0.68**(0.07,6.79) | **0.68**(0.07,7.01) | |
| | $<$ 12h | 37 | 10 | 1[Reference] | 1[Reference] | |
| Weight (g) | | – – – | – – – | **1.00**(0.996,1.006) | **0.999**(0.994,1.005) | – – – – – – – |
| Maternal Age (Years) | | – – – | – – – | **1.02**(0.92,1.14) | **1.01**(0.9,1.13) | – – – – – – – |
| GA (Weeks) | | – – – | – – – | **2**(0.65,6.17) | – – – – – – – | – – – – – – – |

Binary Logistic Regression: (a) simple;(b) adjusted for GA;(c) multivariate with stepwise procedure including GA; *=74 excluding missing values.

O modelo multivariado para os fatores intrínsecos contém apenas a variável Multifetal Gestation, que não é estatisticamente significativa apresentando $p - value$=0.052 (Tabela 2.5c). Os procedimentos do tipo stepwise incluem/retiram variáveis conforme uma significância estabelecida a priori, neste caso 0.05. A inclusão desta variável não significativa, poderá advir de uma adversidade do algoritmo, em que é preferível manter o modelo com a variável Multifetal Gestation que apresentou

um $p - value$ de 0.052 ao modelo nulo. Não será efetuado qualquer tipo de avaliação ao modelo devido à não significância desta variável. Todavia, esta apresenta OR na ordem dos 3 e intervalos de confiança quase excluídos do valor 1. Assim, poderá ser um indício de que esta variável tenda a ser promissora de morbilidade.
Dadas estas circunstâncias, optou-se por efetuar um estudo por bootstrap cujo intuito será compreender a significância desta variável a 5%. Esta abordagem consistiu em 10000 realizações da base de dados com repetição, enfatizando que as amostras geradas são distintas. Foram registados para cada amostra, o *p-value* associado à variável Multifetal Gestation. A figura 2.4 apresenta um histograma com os resultados obtidos, onde no eixo horizontal são apresentados os *p-values* e no eixo vertical a respetiva frequência nas realizações.

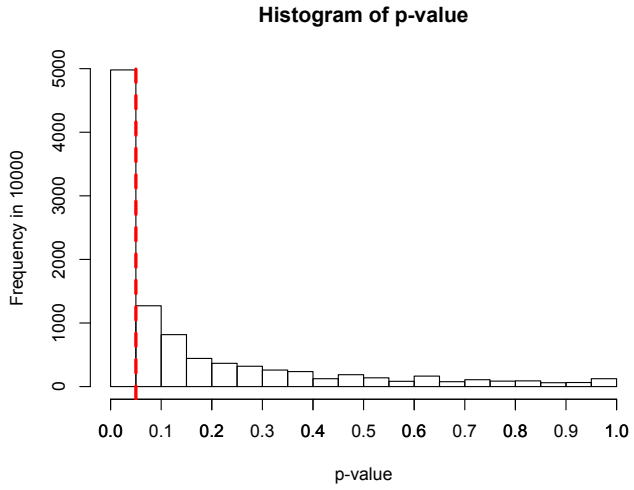



Figura 2.4: Histograma do *p-value* associado à variável Multifetal Gestation, considerando *bootstrap.*

Constatou-se que 50.2% das realizações conduzem a um *p-value* superior a 0.05, sugerindo este resultado que não há evidência estatística de que a variável Multifetal Gestation é um fator de risco significativo considerando uma significância de 5%. Possivelmente, com uma amostra de tamanho maior e mantendo o desvio padrão, Multifetal Gestation já poderia ser considerada fator de risco significativo.

## 2.5 Conclusões

Enquanto que no estudo de mortalidade foi possível a obtenção de fatores de risco e modelos multivariados com significância estatística, embora com impacto limitado quando utilizados para previsão, no estudo de morbilidade, como era de esperar, nenhum fator de risco foi significativo. A obtenção destes resultados poderá ser explicada pelo facto de as variáveis consideradas terem sido recolhidas muito previamente, podendo eventualmente não ter capacidade de explicar o desfecho de mortalidade e morbilidade a longo prazo. Outra limitação poderá redundar da contingência de os possíveis fatores de risco terem sido estudados em subconjuntos diferentes, não permitindo efetuar uma análise multivariada entre todos os candidatos a fatores de risco. Adicionalmente, este estudo tem uma outra limitação subjacente, na medida em que o tamanho da amostra torna-se relativamente pequeno, e consequentemente a amplitude dos intervalos de confiança é bastante grande. Advém desta limitação o facto de algumas variáveis poderem não ser consideradas como

fatores de risco significativos.

# Capítulo 3

# Modelos Preditivos

A circunstância de no capítulo anterior nos termos baseado somente em variáveis recolhidas previamente suscitou-nos alguma inquietação, desde logo por estas poderem não explicar a totalidade dos eventos.
Esta questão e a sua pertinência motivaram então o estudo de modelos preditivos com inclusão de todas as variáveis presentes na base de dados. Desta forma, fatores de risco reportados sistematicamente na literatura como indicadores de morbilidade, associados a problemas motores, cognitivos e respiratórios, como por exemplo IVH, PVL, BPD, HMD terão também oportunidade de constarem na análise.

Neste capítulo serão construídos modelos preditivos de mortalidade e morbilidade baseados em árvores de decisão e regressão logística. Pretende-se, para o efeito, explorar ao máximo toda a informação presente na base de dados, tendo como intuito a identificação de um modelo adequado para cada caso.

Parte deste trabalho foi apresentado nas 4$^{a}$as Jornadas de Iniciação à Investigação Clínica do Centro Hospitalar do Porto e no XX Congresso da Sociedade Portuguesa de Estatística (SPE) (Januário et al., 2012c,b)

## 3.1 Construção de modelos preditivos

Uma das etapas de maior importância para a elaboração de modelos preditivos é a divisão da amostra.

Estes serão desenvolvidos com base na técnica de análise supervisionada (Duda et al., 2001; Hastie et al., 2009), cujo objetivo primordial é prever a classe de um determinado indivíduo, sendo que todo este procedimento é desenvolvido apenas num subconjunto da amostra total, como sugere a figura 3.1.
Dada a informação de inúmeros tratamentos, anomalias e questões fisiológicas de um conjunto de recém nascidos prematuros, incluindo o desfecho destes, que indica mortalidade/sobrevivência ou morbilidade severa/não severa, pretende-se para cada caso, construir um modelo que seja capaz de prever o desfecho (classe) de um novo bebé. Assim, esta abordagem é denominada de "supervisionada", uma vez que para além das variáveis explicativas, a variável resposta (classe) também se encontra presente na orientação do processo de aprendizagem.

No âmbito deste trabalho de investigação, consideremos alguma notação e terminologia adotada. Defina-se $X = (X_1, X_2, ..., X_p)$ como sendo o vetor das $p$ variáveis explicativas e $Y$ a variável resposta binária correspondente às $w_k$ classes, $Y = \{w_0, w_1\}$. Entenda-se que a classe $w_0$ está associada a Alive ou Non Severe e a classe $w_1$ a Dead ou Severe, consoante o estudo de mortalidade e morbilidade (Tabela 1.2). Defina-se ainda $\pi_k$ como sendo as probabilidades *a priori* em cada classe $w_k$, $k = 0, 1$. Idealmente, estas probabilidades deveriam traduzir a prevalência de cada classe na população. No entanto, tais probabilidades são difíceis de se obter, pelo que se considera que as probabilidades *a priori* traduzem a proporção de indivíduos da amostra de treino em cada classe.

Associados à técnica de análise supervisionada, existem diversos métodos que podem ser utilizados para construir um modelo preditivo. Nesta investigação serão abordados apenas dois (Figura 3.1): árvores de classificação e regressão logística.
Embora estes métodos difiram bastante no processo de construção de um modelo, estamos cientes de que a dimensão da amostra é bastante reduzida para o número de variáveis a considerar. Para além disto, considerar um modelo com muitas variáveis poderá reduzir a precisão da previsão e consequentemente, num modelo com poucas variáveis, correr-se-á o risco de se omitir informação relevante. Neste contexto, serão tidos em conta métodos de seleção de variáveis (Hosmer and Lemeshow, 2000; Guyon and Elisseeff, 2003; Hastie et al., 2009), de forma a balancearmos o número de variáveis adequadas de acordo com o tamanho da amostra. É crucial encontrar um modelo que seja capaz de descrever os fenómemos em estudo, através de variáveis verdadeiramente relevantes.

Além da diferença na construção do modelo, as árvores de classificação e a regressão logística são também distintas no resultado que retornam:

- Árvores de classificação: fornecem como resultado um **valor discreto**, obtendo-se diretamente a classe $w_k$, $k = 0, 1$, em que cada indivíduo foi previsto;
- Regressão logística: fornecem como resultado um **valor contínuo**, apresentam-nos uma estimativa da probabilidade *a posteriori*, $p_k$, de um determinado indivíduo pertencer a uma das classes $w_k$, $k = 0, 1$.

Após a construção de um modelo preditivo, é necessário proceder-se à avaliação do mesmo. Esta etapa é já efetuada num subconjunto independente de dados (Figura 3.1). De forma a que seja

possível comparar o desempenho preditivo global dos modelos criados por ambos os métodos, recorrer-se-á a duas métricas: precisão (Overall Accuracy) e AUC (Figura 3.1), sendo que estas têm características diferentes. O cálculo da precisão do modelo tem em consideração o peso das classes, enquanto a AUC (retirada através das curvas ROC) indica se o modelo é capaz de classificar corretamente as classes a prever sem ter em conta a representatividade das mesmas. Ambas serão utilizadas simultaneamente permitindo uma melhor avaliação do modelo preditivo na atribuição de classe a um novo recém nascido prematuro extremo.

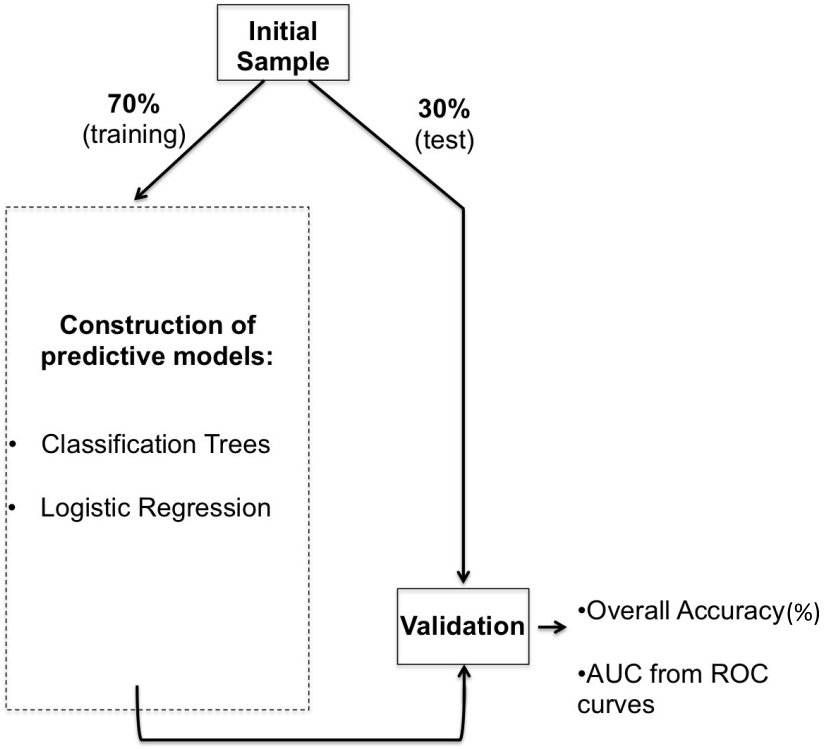



Figura 3.1: Esquema ilustrativo das etapas adotadas neste trabalho para a construção de um modelo baseado em análise supervisionada.

Nas subsecções inerentes a esta secção serão destacadas algumas considerações a ter em conta para a construção dos modelos. Inicialmente, a subsecção 3.1.1 fará alusão à necessidade de preparação da amostra, sendo que nas subsecções 3.1.2 e 3.1.3 serão abordadas a questão do número de variáveis a conter no modelo, bem como estratégias para a redução das mesmas. Posteriormente, a subsecção 3.1.4 abordará com detalhe a avaliação que será efetuada aos modelos de previsão obtidos. Por fim, as restantes secções são vocacionadas para a descrição dos métodos, apresentação de resultados e discussão técnica dos mesmos.

### 3.1.1 Preparação da amostra

O tratamento dos dados e a preparação da amostra é fundamental antes da construção de modelos preditivos. A amostra deve ser explorada cuidadosamente, tentando evitar variáveis mal codificadas e identificando valores em falta ou observações não relevantes para o estudo. Esta etapa é considerada por Soibelman and Kim (2002) como uma das partes mais importantes do processo de construção de modelos sendo, consequentemente, a mais dispendiosa em termos de tempo.

Após a preparação da amostra, é crucial esta ser dividida aleatoriamente em duas amostras independentes:

- **amostra de treino (training sample)**
- **amostra de teste (test sample)**

A amostra de treino é utilizada para construir um modelo de classificação, baseada em toda a informação das observações inclusive as respetivas classes a que pertencem. A amostra de teste permite-nos posteriormente avaliar a precisão do modelo preditivo escolhido. Para que este resultado seja credível é importante que esta amostra de teste não tenha sido utilizada em nenhuma etapa da criação do modelo (Witten et al., 2011).
No entanto, torna-se difícil adotar uma regra sobre como escolher o número de observações em cada uma das amostras, pois esta questão dependerá do número de observações disponível (Hastie et al., 2009).

Adicionalmente, quando se pretende dividir a amostra, é aconselhável que as divisões que dela advêm sejam representativas do problema (Witten et al., 2011). Contudo, quando as amostras de treino e teste são geradas aleatoriamente, não se consegue ter a perceção se esta representa exatamente a variabilidade da população (Soibelman and Kim, 2002).

A figura 3.2 apresenta a preparação e a divisão da amostra em estudo para a construção de modelos preditivos de mortalidade e de morbilidade. Para o caso do modelo de mortalidade, optou-se por excluir da amostra inicial os bebés que nascem mortos (SB), os que morrem na sala de parto (DRD) e aqueles cujo desfecho é desconhecido, pelas mesmas razões já explicitadas no Capítulo 1 (subsecção 1.3.1).

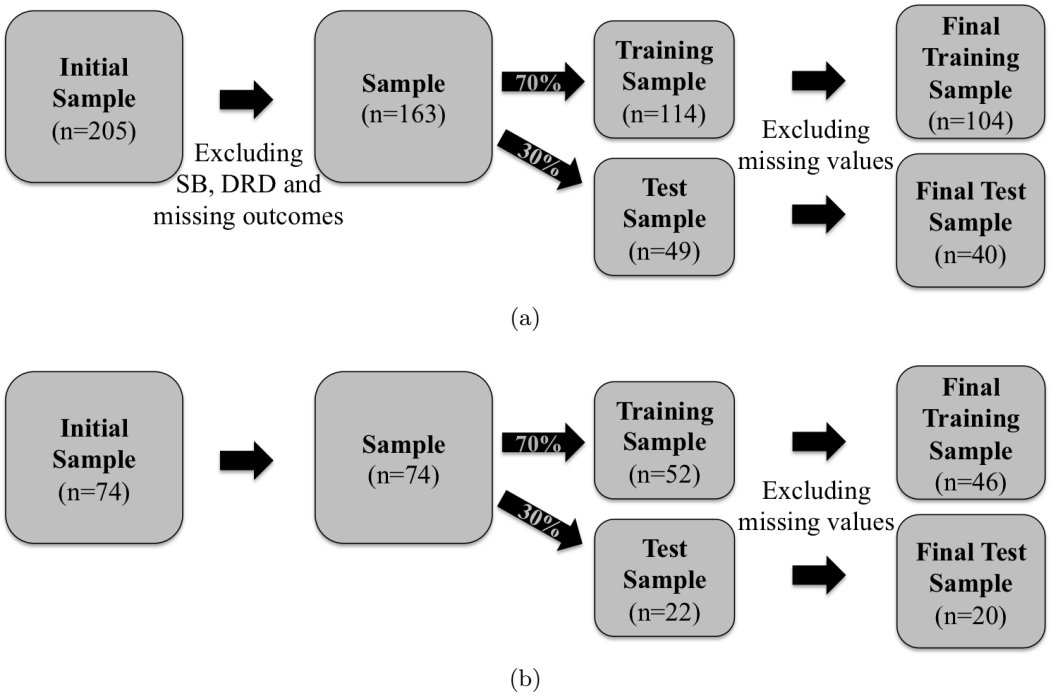



Figura 3.2: Preparação e divisão da amostra inicial para a construção de modelos preditivos de (a) mortalidade e de (b) morbilidade.

De seguida, a amostra de treino foi obtida escolhendo aleatoriamente 70% das observações da amostra total, sendo os restantes 30% constituintes da amostra de teste, à semelhança de um exemplo apresentado em (Soibelman and Kim, 2002).
As amostras de treino e teste obtidas possuem valores em falta, e serão utilizadas no método das árvores de decisão, uma vez que estas conseguem lidar com tais valores. Já o facto de o método de regressão logística não permitir a inclusão de valores desconhecidos, levou a que fossem obtidas novas amostras de treino e teste resultantes da eliminação dos valores em falta das amostras anteriores.

No que se refere ao modelo de morbilidade, não foi necessário uma preparação anterior da amostra, pois esta corresponde aos bebés que sobreviveram ao fim de um ano (subconjunto da amostra de mortalidade). O restante procedimento de divisão da amostra foi equivalente ao de mortalidade.

Destaque-se que neste estudo, as amostras foram obtidas aleatoriamente, não se tendo em conta a representatividade das classes em cada amostra. Assim, esta poderá ser uma limitação do estudo no que concerne à construção de modelos preditivos.

Por vezes, torna-se necessário definir alguns parâmetros do modelo antes que este seja acedido através da amostra de teste.
Quando a amostra é consideravelmente elevada, é frequentemente utilizada a divisão da amostra total em três subconjuntos: treino, validação e teste (Hastie et al., 2009; Witten et al., 2011). O treino serve para construir um modelo de classificação, de acordo com toda a informação disponível; a amostra de validação, como o próprio nome indica, é usada para validar o modelo construído, permitindo otimizar parâmetros do classificador, caso seja necessário; a amostra de teste é utilizada somente para testar o modelo escolhido anteriormente e perceber qual a sua capacidade preditiva. Realce-se que todas estas amostras são independentes e utilizadas apenas nas respetivas etapas, para que quando testado o modelo, a precisão seja confiável.

Caso contrário, quando a amostra é de dimensão reduzida, é preferível considerar apenas duas divisões da amostra: treino e teste. Assim, com esta divisão em duas partes, os modelos são criados através da amostra de treino e caso seja necessário validar algum parâmetro, utilizar-se-á o método de validação cruzada *V-fold*.

**Método de validação cruzada *V-fold***

O método de validação cruzada é frequentemente utilizada quando a dimensão das amostras de treino e teste são consideravelmente pequenas (Witten et al., 2011). Esta técnica é particularmente útil quando necessitamos de definir determinados parâmetros para otimizar o classificador construído através da amostra de treino. Será neste contexto que este procedimento irá ser utilizado neste trabalho.

A essência do método de validação cruzada *V- fold* consiste em dividir aleatoriamente a amostra de treino em $v=1,\ldots,V$ subconjuntos, constituídos sempre que possível pelo mesmo número de elementos. A cada passo, um destes $V$ subconjuntos, é reservado para o teste, enquanto os restantes *V-1* são utilizados como sendo a nova amostra de treino para a construção do modelo (Figura 3.3). Desta forma, são sempre considerados $\frac{V-1}{V}$ % dos dados para treino e os restantes para teste. Este processo de construção de modelos é então repetido $V$ vezes, considerando em cada etapa um subconjunto de teste e de treino diferentes. Assim, são construídos $V$ modelos diferentes

que serão testados em subconjuntos independentes dos de treino, em cada etapa $v$. Finalmente, obtêm-se $V$ estimativas de erros calculadas com recurso ao subconjunto de teste destinado a cada etapa, produzindo-se uma estimativa de erro global correspondente à média aritmética dos $V$ subconjuntos.

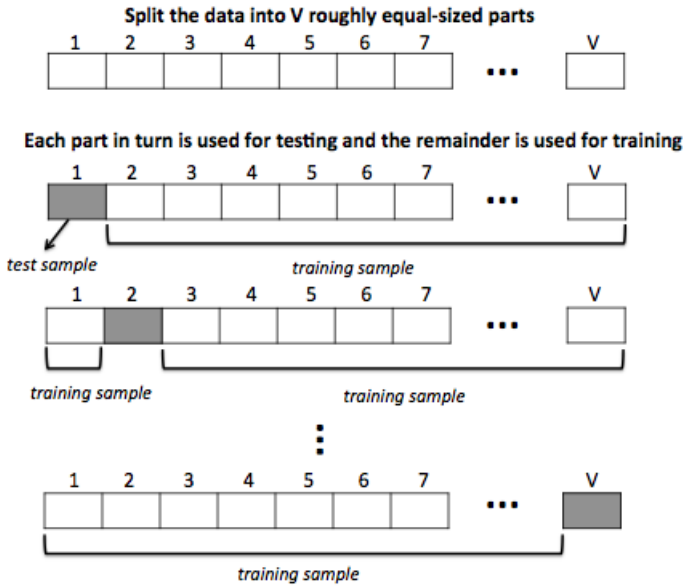



Figura 3.3: Exemplificação do método de validação cruzada *V-fold*, adotado neste trabalho. Figura inspirada em Hastie et al. (2009).

A questão de quantos subconjuntos devem ser considerados é bastante frequente, existindo estudos relativos de qual será o valor de $V$ mais adequado. Neste trabalho de investigação adotar-se-á *V=10*, uma vez que se trata do valor mais comum recomendado pela literatura (Witten et al., 2011; Hastie et al., 2009; Kohavi, 1995). Assim, a cada passo serão usados para a construção de cada modelo 90 % dos dados para treino e os restantes 10% para teste (Figura 3.3).

### 3.1.2 Número de variáveis no modelo

Em concordância com a secção anterior, a amostra inicial foi dividida de acordo com o intuito do estudo, sendo que todas as amostras que dela derivam são constituídas por 23 variáveis categóricas e 3 contínuas (consultar Tabela 1.3).
Se a dimensão da amostra inicial deste estudo, quer para o caso da mortalidade ($n$=205) quer para o caso de morbilidade ($n = 74$) já é considerada reduzida, as amostras de treino a partir das quais se irão construir modelos preditivos é ainda mais pequena (Figura 3.2). Neste contexto, a dimensão da amostra de treino restringe, sem dúvida, o número de variáveis a incluir num modelo multivariado. De facto, conclui-se que o número de observações torna-se demasiado pequeno para o número de variáveis em estudo.

Para contornar a situação anterior, foram primeiramente utilizados em regressão logística os métodos mais comuns de seleção de variáveis: *stepwise*, *backward elimination* e *forward selection* (Hastie et al., 2009; Hosmer and Lemeshow, 2000), com o intuito de se obter um modelo com um número de variáveis mais reduzido. Todavia, ambos os métodos apresentaram um problema

comum: não conseguiram encontrar um modelo que convergisse.

Posteriormente, foi efetuada uma pesquisa bibliográfica cujo objetivo principal incidiu na procura do número adequado de variáveis a incluir num modelo multivariado, de acordo com a dimensão da amostra. Segundo uma recomendação da literatura, devem ser considerados 10 eventos por variável (Peduzzi et al., 1996). No caso da regressão logística, os autores explicitam que o número de eventos a considerar, para aplicação desta regra, deverá ser o menor valor do outcome binário (Dead e Alive ou Severe e Non Severe).
Contudo, estudos mais recentes (Vittinghoff and McCulloch, 2007) indicam que esta *"regra de ouro"* é demasiado conservativa e pode de facto ser relaxada. A conclusão levada cabo por estes autores baseia-se em experiências que estes efetuaram, fazendo variar o número de eventos. Não obstante, estes indicam que quando se consideram menos de 10 eventos, os resultados obtidos deverão ser interpretados cautelosamente. Para o caso dos modelos de regressão logística, determine-se o número de variáveis a considerar:

- **mortalidade**

A amostra de treino final para a construção do modelo de mortalidade apresenta uma dimensão de $n = 104$ (Figura 3.2a), em que 55 dos bebés têm como outcome Dead e os restantes 49 Alive. Temos então que os modelos deverão incluir no máximo: $\frac{49}{10} \approx 5$ variáveis.

- **morbilidade**

Neste caso, a amostra de treino possui um total de 46 observações (Figura 3.2b), tendo 35 bebés o outcome Non Severe e os 11 restantes, Severe. De acordo com Peduzzi et al. (1996), o modelo de morbilidade deverá ter no máximo $\frac{11}{10} \approx 1$ variável.

Face à potencialidade desta abordagem nos modelos de regressão logística, decidiu-se seguir uma linha orientadora equivalente nas árvores de classificação. A aproximação desta regra às árvores de classificação será efetuada com base num critério de paragem do crescimento da árvore, tendo em conta o número de observações a conter num determinado nó. Para se obter um modelo de árvores de classificação de tamanho comparável ao de regressão logística, tem-se que:

- **mortalidade**

A amostra de treino para o modelo de mortalidade possui 114 observações (Figura 3.2a). Considerando-se que o modelo deverá conter no máximo as tais 5 variáveis, tem-se que o número de observações num nó deverá ser no mínimo $\frac{114}{5} \approx 20$ observações.

- **morbilidade**

A amostra de treino para o modelo de morbilidade possui 52 observações (Figura 3.2b). Considerando 5 como majorante do número de variáveis da árvore de morbilidade, tem-se que um nó deverá ter no mínimo $\frac{52}{5} \approx 10$ observações.

A tabela 3.1 sumaria os cálculos efetuados para o caso de mortalidade e morbilidade, tendo em conta os métodos a utilizar.

Tabela 3.1: Número adequado de variáveis para modelos de regressão logística e número de observações num nó para árvores de classificação, de acordo com o tamanho da amostra.

| | **Mortality** | | **Morbidity** | |
|---|---|---|---|---|
| | Logistic Regression | Classification Tree | Logistic Regression | Classification Tree |
| **# variables** | 5 | − − − | 1 | − − − |
| **# observations in a node** | − − − | 20 | − − − | 10 |

### 3.1.3 Associação entre variáveis

Visto não ser possível aumentar o número de observações, a única alternativa que resta é tentar reduzir o número de variáveis. Assim, torna-se essencial continuar a procurar um subconjunto adequado das 26 variáveis disponíveis. É fulcral relembrar que destas 26 variáveis apenas 3 são contínuas (Weight, GA, Maternal age), sendo as restantes categóricas nominais.

Das 3 variáveis contínuas, apenas serão consideradas diretamente para os modelos de mortalidade e morbilidade, o peso e a idade gestacional, devido à relevância das mesmas na literatura. Adicionalmente, estas variáveis foram também previamente identificadas como fatores de risco (Capítulo 2).
Para as restantes 23 variáveis, uma primeira etapa, passa então por construir uma matriz de associação entre duas variáveis categóricas, por forma a identificar pares de variáveis muito relacionadas. Numa segunda etapa, após a identificação de conjuntos de variáveis muito associadas, podem-se efetuar as seguintes escolhas:

- de cada um dos subconjuntos formados, optar por se escolher apenas a mais associada com o evento (mortalidade/morbilidade). As restantes, que não conseguiram formar nenhum subconjunto seriam então utilizadas independentemente;
- mergir duas variáveis que estejam muito associadas, colocando toda a informação referente às duas numa nova variável.

Para o procedimento de construção da matriz de associação, é necessário escolher uma medida adequada para variáveis categóricas nominais.

Segundo a literatura (Liebetrau, 1983; Goodman and Kruskal, 1954), as medidas de associação mais comuns para este tipo de variáveis, derivam da estatística do qui-quadrado, $\chi^2$ , na qual se baseiam os testes de independência:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.1}$$

onde $O_{ij}$ e $E_{ij}$ representam a frequência observada e a frequência esperada na célula $ij$, respetivamente. Whitford (2005) faz uma revisão de diversas medidas de associação, tendo em conta o tipo de variáveis.
Para lidar com variáveis categóricas nominais, as medidas de associação mais frequentes são as seguintes:

- **Coeficiente de contingência ($C$)**

- **Coeficiente de Phi ($\phi$)**
- **Coeficiente de Cramer ($V$)**

Este coeficientes têm uma interpretação equivalente ao coeficiente de correlação de Pearson, utilizada para exibir a correlação entre duas variáveis quantitativas. No entanto, os coeficientes indicados para medir a "força"entre variáveis categóricas nominais assumem apenas valores positivos e geralmente compreendidos entre 0 e 1, não havendo associação negativa entre variáveis. Uma medida de associação com o valor 1 indica que duas variáveis estão fortemente associadas.

**Coeficiente de contingência**
O coeficiente de contingência pode ser calculado para qualquer tabela de contingência, sendo obtido da seguinte forma:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{3.2}$$

onde $\chi^2$ representa a estatística do qui-quadrado e $n$ o tamanho da amostra. Contudo, o valor máximo deste coeficiente varia de acordo com o tamanho da amostra, não permitindo uma comparação razoável entre variáveis com tabelas de contingência diferentes.

**Coeficiente de Phi**
Este coeficiente pode ser utilizado apenas para tabelas de contingência 2×2, ou seja, só é aplicável para variáveis com duas categorias. O cálculo a efetuar é o seguinte:

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{3.3}$$

onde $X^2$ refere-se à estatística do qui-quadrado e $n$ a dimensão da amostra.

**Coeficiente de Cramer V**
Do inglês *"Cramer's V coefficient"*, este coeficiente é também aplicável a qualquer tipo de tabela de contingência e surge com o intuito de contornar a limitação do coeficiente de contingência, $C$, variando num intervalo entre 0 e 1. A fórmula é novamente baseada na estatística do qui-quadrado mas também ajustada de acordo com o número de linhas e colunas presentes na tabela de contingência

$$V = \sqrt{\frac{\chi^2}{n[min(r,c) - 1]}} \tag{3.4}$$

onde $\chi^2$ é a estatística usual do qui-quadrado, $n$ o tamanho da amostra e $(r, c)$ o número de (linhas, colunas) da tabela de contingência.
Note-se que caso a tabela de contingência seja 2×2, $min(r, c) - 1$ é 1. Assim, o coeficiente de $\phi$ é um caso particular do coeficiente de Cramer V, quando se trata de associação entre variáveis dicotómicas.

Pela análise antecedente, e tendo em conta que algumas das variáveis categóricas nominais em estudo possuem mais de duas categorias, o **coeficiente de Cramer V** traduz-se na medida de associação mais adequada. Neste sentido, a associação entre pares de variáveis categóricas $X_i$ e $X_j$ é apresentada na respetiva célula da matriz, onde $V_{i,j} = 1$ indica forte associação entre as variáveis.

Como é evidente, a identificação de subgrupos de variáveis muito associadas através de uma matriz de associação **V**, pode tornar-se um processo bastante difícil, na medida em que é necessário

analisar cada linha e as respetivas colunas com muito detalhe.

Foi então que surgiu a ideia de recorrer à análise classificatória hierárquica, para de forma intuitiva aceder-se ao agrupamento sucessivo de variáveis e obter os subgrupos (*clusters*) de variáveis mais associadas, caso existam. Este tipo de classificação, exige que seja definida uma medida de dissemelhança entre os pares de variáveis, que como o próprio nome indica, espelhe as diferenças entre tais variáveis (Hastie et al., 2009; Witten et al., 2011; Ripley, 1996; Duda et al., 2001).
A ideia da utilização deste método de análise hierárquica, será tentar reduzir o número de variáveis. Posteriormente, através de uma pesquisa bibliográfica sobre este assunto, detetamos que a análise de clusters tem vindo a ser usada em algumas investigações que procuram selecionar um subconjunto de variáveis (Guyon and Elisseeff, 2003), embora seja mais usual a análise classificatória através do algoritmo *K-means*.

Neste contexto, a análise hierárquica foi efetuada considerando como dissemelhança a matriz **1-V**, cuja diagonal é constituída por zeros, indicando não dissemelhança quando comparadas as próprias variáveis. Assim, variáveis muito associadas terão uma dissemelhança próxima de zero, sendo agrupadas em níveis mais baixos.
Para o agrupamento das inúmeras variáveis foram considerados diversos tipos de ligação (Witten et al., 2011), tais como: *índice do mínimo*, *índice do máximo*, *índice da média*, *índice de Ward* e *índice do centro de gravidade*.
Posteriormente, com o propósito de se verificar qual o índice de ligação que permite uma hierarquia mais apropriada do problema em estudo, estes foram comparados através do coeficiente de correlação cofenético, que visa ser uma medida equivalente a uma correlação entre variáveis e, através da deformação delta (Hastie et al., 2009). A representação hierárquica é tanto melhor quanto menor for a deformação delta e, consequentemente, quanto maior for o coeficiente de correlação cofenético.

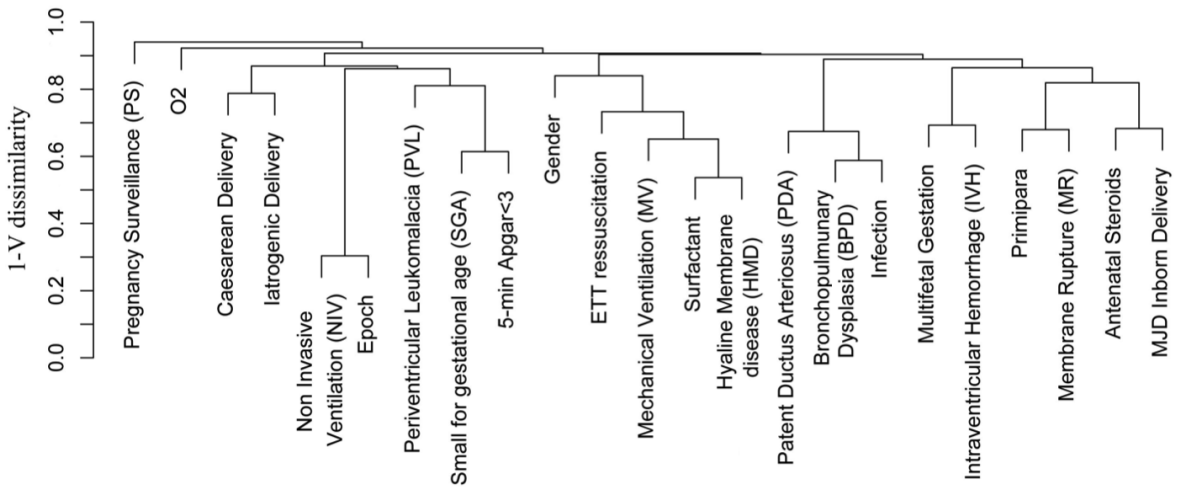
A tabela 3.2 sumaria os valores do coeficiente de correlação cofenético e da deformação delta para a amostra de treino de mortalidade e morbilidade, considerando os diversos índices. Em ambos os casos, o *índice de ligação da média* mostrou ser o mais adequado, apresentando simultaneamente uma menor deformação e um maior coeficiente de correlação cofenético.

Tabela 3.2: Valores do coeficiente de correlação cofenético e da deformação delta de acordo com os índices do *mínimo*, *do máximo*, *da média*, *de Ward* e *do centro de gravidade* para o caso de mortalidade (preto) e morbilidade (azul), respetivamente.

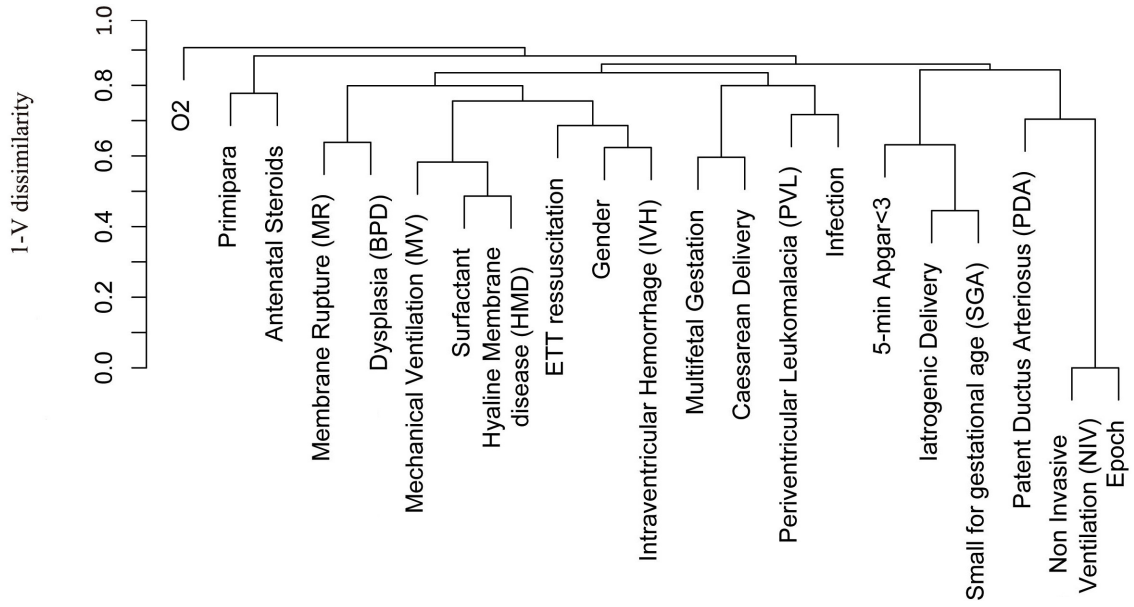
| | **Índice** | | | | |
|---|---|---|---|---|---|
| | *mínimo* | *máximo* | *média* | *Ward* | *centro de gravidade* |
| Coeficiente de Correlação Cofenético | 0.63;0.61 | 0.69;0.67 | **0.73**;**0.74** | 0.50;0.44 | 0.22;0.55 |
| Deformação Delta | 5.35;7.08 | 3.04;4.69 | **1.07**;**1.42** | 58.70;61.27 | 39.93; 28.73 |

Uma forma prática e visualmente intuitiva para identificar subconjuntos de variáveis muito associadas passa por representar a hierarquia graficamente, através do denominado *dendrograma*. Este tipo de gráfico permite de forma sistemática explorar se existe uma estrutura, ou seja, apresenta os agrupamentos das variáveis de acordo com a sua dissemelhança.
A figura 3.4 apresenta os dendrogramas de mortalidade e morbilidade obtidos pela medida de dissemelhança $1 - V$ de acordo com *índice da média*. Em ambos os casos, as variáveis mais

associadas mostraram ser a Epoch e NIV, apresentando mesmo dissemelhança nula no estudo de morbilidade (Figura 3.4b). No caso da mortalidade, a variável PS é a última a ser incluída na ligação (Figura 3.4a), enquanto que no caso da morbilidade a última variável a ser agrupada diz respeito a O2 (Figura 3.4b).




(a) Mortality




(b) Morbidity

Figura 3.4: Dendrogramas ilustrando o agrupamento das variáveis do estudo, baseados no índice da média e dissemelhança **1-V**, onde **V** representa o **coeficiente de Cramer V** (equação (3.4)).

A figura 3.4 mostra que de facto as variáveis são agrupadas com valores de ligação muito elevados, indicando que qualquer subconjunto que possa ser considerado, irá ter variáveis muito heterogéneas. Esta situação de semelhança era de esperar ser observada em ambos os casos, uma vez que as observações da amostra de morbilidade são, na realidade, um subconjunto das observações de mortalidade (referem-se aos bebés que sobreviveram). No entanto, considerou-se necessário efetuar

a mesma análise pois sendo um subconjunto de observações, pode traduzir algumas associações com maior evidência estatística, como foi o caso.
Torna-se importante realçar que o dendrograma associado a mortalidade (Figura 3.4a) apresenta as 23 variáveis categóricas da amostra de treino. Não obstante, a variável BPD foi posteriormente retirada do estudo de mortalidade, devido à sua inadequada codificação. Também uma inspeção cuidada da amostra de treino de morbilidade antes de se efetuar qualquer tipo de análise, levou a que duas variáveis, PS e MJD Inborn Delivery fossem excluídas do estudo, não aparecendo portanto no respetivo dendrograma (Figura 3.4b). Esta eliminação, é justificada pelo facto de estas variáveis assumirem sempre o mesmo valor para todas as observações, indicando que todos os bebés tiveram uma gravidez vigiada e nasceram na MJD. Deste modo, estas variáveis não trariam qualquer benefício de serem incluídas na análise, uma vez que o seu comportamento não difere.

A forte associação entre as variáveis Epoch e NIV em ambos os casos, merece especial destaque, motivando uma análise entre as duas variáveis. A tabela 3.3 apresenta tal associação entre as variáveis, observando-se que entre os anos 2000 a 2006, a ventilação não invasiva (NIV) não era utilizada, enquanto que a partir da época de 2007, NIV começou a ser uma prática clínica, sendo que todos os bebés que sobreviveram (amostra de mortalidade) usufruíram de tal tratamento. Este resultado espelha nitidamente as mudanças e atitudes na prática clínica, justificadas pela equipa médica da MJD.

Tabela 3.3: Associação entre a variável NIV e Epoch na amostra de treino de mortalidade (preto) e morbilidade (azul), respetivamente.

| | Epoch | | | |
|---|---|---|---|---|
| | (2000-2002) | (2003-2006) | (2007-2009) | Total |
| NIV=No | 36;18 | 51;20 | 8;0 | 95;38 |
| NIV=Yes | 0;0 | 0;0 | 9;8 | 9;8 |
| Total | 36;18 | 51;20 | 17;8 | 104;46 |

### 3.1.4 Desempenho do modelo

Um modelo preditivo tem como objetivo primordial desempenhar previsões favoráveis relativamente à variável que se pretende prever.
No contexto deste trabalho de investigação, as variáveis a predizer assumem apenas dois valores discretos (classes) possíveis:

- **mortalidade**: Dead ($w_1$) ou Alive ($w_0$)
- **morbilidade**: Severe ($w_1$) ou Non Severe ($w_0$)

Após a construção de um modelo de previsão na amostra de treino, pretende-se então avaliar o seu desempenho na amostra de teste. O resultado da previsão de classe a cada indivíduo é dependente do tipo de modelos de classificação utilizados. Podem-se distinguir dois tipos de modelos:

- os que produzem como resultado um **valor discreto**, indicando diretamente a classe prevista, como é o caso das *árvores de classificação*.
- os que apresentam como resultado um **valor contínuo** (i.e, uma estimativa da probabilidade *a posteriori*, $p_k(k = 0, 1)$, de um determinado indivíduo pertencer a uma das classes (geralmente a $w_1$) para o qual diversos *pontos de corte* da $p_k$ poderão ser aplicados para

determinar o valor discreto da classe. Nesta situação, verifica-se que para cada valor de corte poderão resultar diferentes previsões (Fawcett, 2006; Prati et al., 2008). É exemplo desta situação o modelo de *regressão logística*.

Uma forma inata de avaliar um modelo de classificação, passa por apresentar uma tabela de dupla entrada, permitindo aceder ao cruzamento entre as classes previstas e as classes observadas na amostra de teste. Esta tabela é na maioria das vezes, intitulada de *matriz de confusão*. A tabela 3.4 ilustra-nos uma *matriz de confusão* generalizada para o caso de duas classes, designadas por $w_1$ e $w_0$, respetivamente.

Tabela 3.4: Matriz de confusão, onde $w_1$ representa a classe do evento em estudo que pretendemos prever (Dead ou Severe) e $w_0$ a classe contrária.

| | | Observed | | Overall |
|---|---|---|---|---|
| | | $w_1$ | $w_0$ | Accuracy(%) |
| Predicted | $w_1$ | True Positive ($TP$) | False Positive ($FP$) | |
| | $w_0$ | False Negative ($FN$) | True Negative ($TN$) | |
| | | $\frac{TP}{\#w_1} \times 100$ | $\frac{TN}{\#w_0} \times 100$ | $\frac{TP+TN}{\#w_1+\#w_0} \times 100$ |

É notório que tendo em conta um modelo de previsão e uma determinada observação, existem quatro desfechos possíveis (Tabela 3.4):

- Verdadeiro Positivo ($TP$): uma observação da classe $w_1$ é prevista como $w_1$;
- Falso Positivo ($FP$): uma observação da classe $w_0$ é prevista como $w_1$;
- Falso Negativo ($FN$): uma observação da classe $w_1$ é prevista como $w_0$;
- Verdadeiro Negativo ($TN$): uma observação da classe $w_0$ é prevista como $w_0$.

Intuitivamente, quando se pretende construir um classificador, a meta é tentar minimizar os dois tipos de erros possíveis, isto é, minimizar os $FP$ e os $FN$. Equivalentemente, pretende-se maximizar as classificações corretas $TP$ e $TN$. Neste sentido, são frequentemente utilizadas duas medidas de desempenho (*Sensitivity* e *Specificity*) que permitem avaliar a capacidade preditiva do modelo, através da percentagem de classificações corretas em cada classe:

$$Sensitivity(\%) = \frac{TP}{\#w_1} \times 100 \quad e \quad Specificity(\%) = \frac{TN}{\#w_0} \times 100 \tag{3.5}$$

Contudo, para a avaliação global de um modelo de previsão é comum utilizar-se como medida a precisão do modelo, que considera simultaneamente $TP$ e $TN$:

$$Overall\ Accuracy(\%) = \frac{TP + TN}{\#w_1 + \#w_0} \times 100 \tag{3.6}$$

Apesar da avaliação do modelo ser baseada essencialmente na matriz de confusão é, no entanto, necessário ter em atenção se as medidas de desempenho consideradas são as mais adequadas. Batista et al. (2004) e Prati et al. (2008) referem que o erro de classificação e a *Overall Accuracy* do modelo poderão induzir a conclusões erradas, nomeadamente quando as probabilidades *a priori*, $\pi_k$, $k = 0, 1$, em cada classe são muito distintas, isto é, as classes não são homogéneas. Este facto

tem como consequência a classificação das observações na classe maioritária.

Uma ferramenta útil e bastante utilizada nos últimos tempos para a avaliação de classificadores são as denominadas curvas ROC (do inglês *"Receiver Operating Characteristics"*). O espaço da curva ROC é caracterizado como sendo bidimensional, onde são apresentados no eixo das ordenadas, valores de *Sensitivity* (taxa de *TP*) e no eixo das abcissas 1-*Specificity* (taxa de *FP*) (Figura 3.5). No âmbito da teoria de deteção de sinais, corresponderia a ilustrar graficamente a taxa de detetar corretamente um sinal verdadeiro (*Sensitivity*) e um sinal falso (*1-Specificity*).

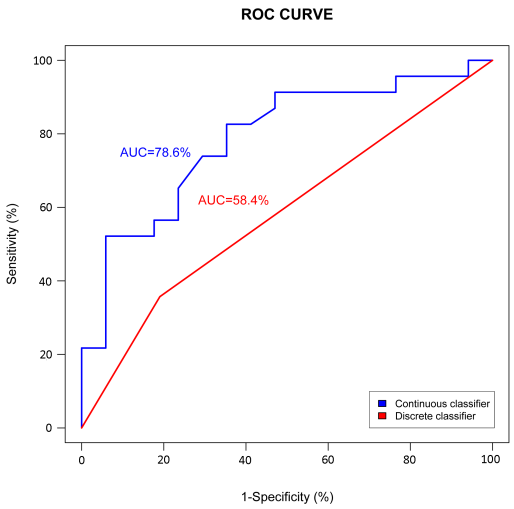



Figura 3.5: Exemplo do espaço ROC para um classificador discreto e um contínuo.

Note-se que a curva ROC é um pouco diferente para classificadores **discretos** e classificadores **contínuos**. Como já mencionado, no caso do modelo de classificação retornar diretamente a classe do objeto, produz-se apenas um par (*1-Specificity*, *Sensitivity*) que corresponderá a um único ponto no espaço ROC, uma vez que existe apenas uma *matriz de confusão*. Caso contrário, quando o modelo retorna um **valor contínuo**, cada *valor de corte* produz um par (*1-Specificity*, *Sensitivity*), dando origem a diferentes pontos no espaço ROC. Assim sendo, dois pontos consecutivos são unidos através de uma linha, originando a curva ROC.
Em suma, a curva ROC permite compreender a variação da *Sensitivity* de acordo com *1-Specificity* (Obuchowskl, 2003), para vários valores de corte. Naturalmente, um bom modelo será o que apresentar maior *Sensitivity* e menor *1-Specificity*. Assim, o modelo é tanto melhor quanto mais próximo se encontrar do canto superior esquerdo. No caso do exemplo retratado na figura 3.5 o melhor modelo é o classificador contínuo. Note-se ainda que no caso de um classificador contínuo, uma possível escolha para um *ponto de corte* ótimo será adotar aquele cuja discretização se encontra mais próxima do ponto (*1-Specificity*,*Sensitivity*)=(0%,100%).

Uma das medidas mais comuns para avaliação e comparação de modelos de classificação através de curvas ROC é a designada área abaixo da curva, AUC (do inglês *"Area under the curve")*.
Percentualmente, a AUC toma valores compreendidos entre 0 e 100 , permitindo avaliar a capacidade de discriminação do modelo. Quanto maior for o valor de AUC mais poder discriminante tem o modelo, pois a *Sensitivity* aproxima-se cada vez mais de 100% e a taxa de falsos positivos (*1-Specificity*) centra-se próximo de 0%.
Concetualmente, Hosmer and Lemeshow (2000) assumem como regra geral:

- AUC=50% : sugere que não há discriminação entre classes;
- 70% $\leq$AUC<80% : considera-se uma discriminação aceitável;
- 80% $\leq$AUC<90% : considera-se uma discriminação excelente.

De forma simplificada, a AUC de um modelo preditivo é equivalente a calcular a proporção de vezes que um objeto da classe $w_1$ tem uma probabilidade superior ao da classe $w_0$. No entanto, existem outras interpretações da AUC, sendo esta também equivalente à estatística de Wilcoxon e correlacionada com o índice de Gini (Fawcett, 2006; Lasko et al., 2005).

As curvas ROC foram introduzidas inicialmente na teoria de deteção de sinais, com o intuito de caraterizar a receção de um sinal na presença de ruído (Fawcett, 2006; Prati et al., 2008; Witten et al., 2011). Este tipo de gráfico tem sido utilizado em inúmeras áreas (economia, finanças, psicologia) mas essencialmente na área da medicina. Nesta área, as curvas ROC são usadas para estudar diversos testes clínicos, como por exemplo: presença ou ausência de uma determinada doença (Obuchowskl, 2003; Lasko et al., 2005) e respetiva capacidade discriminante associada ao classificador, através da AUC. Estudos equivalentes ao desta investigação, utilizam também curvas ROC e AUC para aceder aos desfechos de mortalidade e morbilidade em recém nascidos prematuros extremos (Tyson et al., 2008).

## 3.2 Modelos baseados em árvores de decisão

Nesta secção serão apenas abordadas as árvores de classificação, uma vez que as variáveis a predizer são nominais (Tabela 1.2).

A estrutura de uma árvore tem como ponto de partida uma *raiz*, sendo esta exibida na parte superior da árvore, como é possível visualizar na figura 3.6. A *raiz* é dividida e conectada com os respetivos *nós* através de ligações (*ramos*). Em cada *nó* está presente um teste que dará origem por sua vez a outro nó (através de ligações sucessivas) ou a *nós terminais* (*folha*) caso não seja possível proceder a mais divisões e, neste caso, decide-se qual a classe a que determinado objeto irá pertencer. Note-se que a raiz divide a amostra treino completa e, cada decisão sucessiva de um nó divide um subconjunto adequado dos dados de treino. A cada decisão tomada num nó denomina-se de *divisão* (Duda et al., 2001).

A figura 3.6 permite-nos representar uma progressão da análise de dados com o intuito de explicitar uma determinada tarefa de previsão. Sublinhe-se que a classificação de um objeto consiste apenas em seguir um trajeto definido pelos sucessivos testes da árvore até que seja encontrado um nó terminal que lhe atribuirá a classe.

As árvores de decisão remontam aos anos 60, com especial referência aos trabalhos desenvolvidos por Morgan and Sonquist (1963) na área das ciências sociais. Mais tarde, na década de 80, com o advento do livro de Breiman et al. (1984) surge um novo algoritmo denominado de CART (do inglês, *"Classification And Regression Trees"*). Esta publicação traduziu-se numa das principais ferramentas para a construção de árvores de decisão, sendo ainda hoje um dos algoritmos mais utilizados. Posteriormente, viria a merecer especial destaque o algoritmo ID3 proposto por Quinlan (1986) e que serviu também de base para um novo algoritmo proposto pelo mesmo autor anos mais tarde, C4.5 (Quinlan, 1993). Embora haja diferenças na aplicação dos algoritmos anteriores, ambos se baseiam no mesmo princípio: "dividir para conquistar".

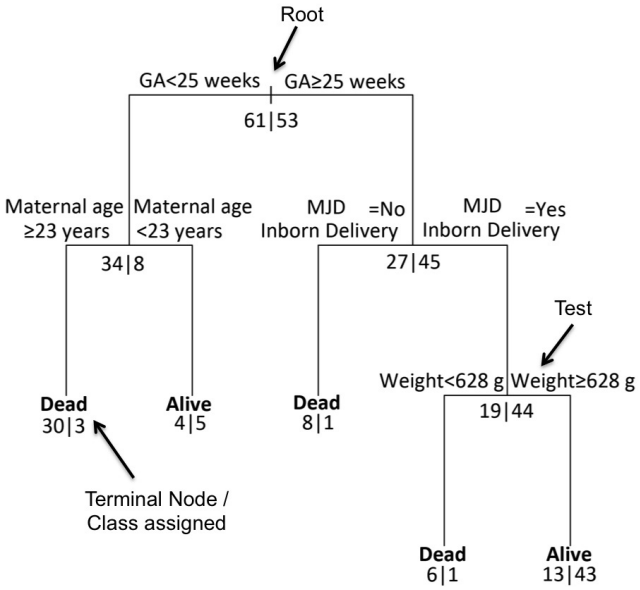



Figura 3.6: Exemplo de uma árvore de classificação. Por exemplo, um bebé que apresente GA$\geq$ 25 weeks, MJD Inborn Delivery= Yes e Weight<628 g é classificado como morto. Neste trabalho adiciona-se a cada nó terminal, #**Dead** | #**Alive** da amostra de treino.

Assim, as árvores de decisão apresentam-se como sendo uma estrutura hierárquica tal como uma árvore, desenvolvem-se da raiz para as folhas, na qual diversas condições são testadas sequencialmente, isto é, existe uma segmentação do problema em vários subproblemas de menores dimensões até que uma determinada solução seja aceite.
A metodologia base de "dividir para conquistar"é utilizada em todos os algoritmos baseados em árvores de classificação. Contudo, estes apresentam inúmeras diferenças na construção da árvore, nomeadamente a sua estrutura, o critério adotado para segmentar um nó, a técnica de poda e também a forma de lidar com os valores em falta (Kohavi and Quinlan, 1999). Neste estudo, apenas será abordado o algoritmo de CART, uma vez que este visa ser bastante influente no contexto da estatística e das máquinas de aprendizagem. Uma análise detalhada dos algoritmos ID3 e C4.5 pode ser consultada no trabalho de Fonseca (1994) bem como na bibliografia referida destes algoritmos.

A figura 3.7 ilustra os passos essenciais na construção de uma árvore de classificação, particularizando-se ao algoritmo de CART. A presente secção será estruturada em diversas subsecções nas quais as três primeiras serão referentes às etapas principais destacadas na figura anterior:

**(1) Binary Recursive Partitioning (subsecção 3.2.1)**
Abordar-se-á o método de particionamento recursivo binário adotado pelo algoritmo de CART bem como a escolha do melhor atributo que deverá segmentar o nó.

**(2) Stopping Criterion (subsecção 3.2.2)**
Serão apresentados diversos critérios de paragem que poderão ser utilizados no processo de construção da árvore.

**(3) Prune (subsecção 3.2.3)**
As duas etapas anteriores permitem obter uma árvore máxima. Nesta subsecção apresentam-se

ideias gerais sobre o algoritmo de poda, cujo objetivo é a criação de uma árvore mais simplificada, denominada de árvore podada.

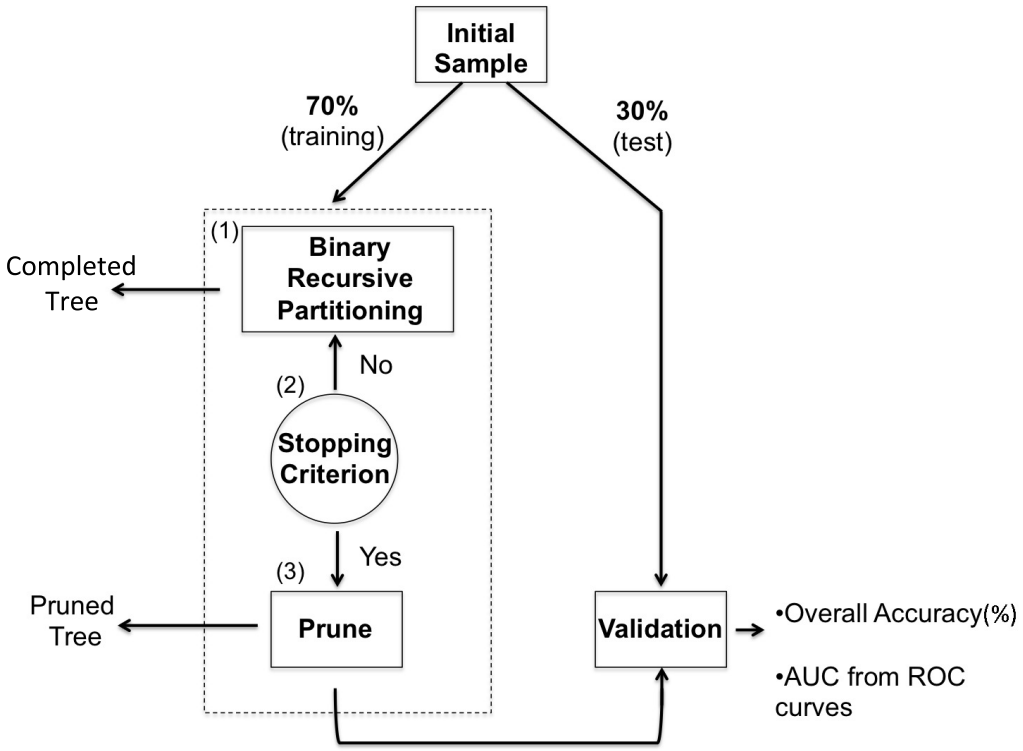



Figura 3.7: Principais etapas na construção de uma árvore de decisão.

Posteriormente, a subsecção 3.2.4 reportar-se-á às estratégias para lidar com os valores em falta, sendo esta uma das principais vantagens das árvores. Na subsecção 3.2.5 são introduzidos modelos considerando custos de má classificação. Por fim, as árvores obtidas através da amostra de treino capazes de prever mortalidade de recém nascidos prematuros ao um ano (morto/vivo), bem como a morbilidade dos mesmos tendo em conta o seu desenvolvimento psicomotor (severos/não severos) serão validadas nos dados de teste e comparadas através da precisão global e de curvas ROC, subsecção 3.2.6.
É de salientar que os exemplos fornecidos em todas as subsecções teóricas serão referentes aos resultados obtidos para o modelo de mortalidade, de forma a ilustrar a descrição teórica.

### 3.2.1 Particionamento Recursivo Binário - construção da árvore

O algoritmo de CART é baseado na técnica *"Binary Recursive Partitioning"*, dado que o resultado obtido é sempre uma árvore binária que pode ser percorrida hierarquicamente respondendo apenas a questões do tipo "sim/não". O termo **binary** implica que cada subconjunto de observações representado por um *nó* $t$ nas árvores de classificação tem de ser segmentado apenas em outros dois *nós*, $t_l$ e $t_r$, dando origem a dois novos subconjuntos. Nesta situação, o nó que foi segmentado é apelidado de *nó pai* e os que dele derivam são designados por *nós filho*. Já o termo **recursive** indica-nos que este processo de segmentação/divisão de nós pode ser repetido sucessivamente a cada um dos subconjuntos gerados até que não sejam possíveis mais divisões. Assim, cada *nó pai* pode originar dois *nós filho* e estes por sua vez podem ser segmentados e originarem ainda outros dois *nós filho*. É de salientar que esta divisão binária recursiva desenvolve-se sempre da raiz

para as folhas. A expressão **partitioning** refere-se ao facto de o conjunto de dados de treino ser particionado em subconjuntos mais pequenos.

A figura 3.8 ilustra a árvore máxima obtida para o modelo de mortalidade em recém nascidos prematuros extremos. Para uma melhor compreensão serão analisados alguns pormenores descritos anteriormente.
Da análise da figura 3.8 depreende-se que cada nó apenas possui divisões binárias e que estes podem ser divididos sucessivamente, como reportado no texto precedente. Para além disto, cada nó responde a questões simples do tipo "sim/não". A figura 3.8a representa uma possível estrutura de uma árvore de classificação em que cada nó contém uma pergunta. Por exemplo, na *raiz* coloca-se a seguinte questão *"o bebé tem GA<25 weeks?"*, e, caso obtivesse resposta verdadeira seguiria o ramo da esquerda representado pela resposta "Yes". Uma representação alternativa a esta e que será utilizada neste trabalho é a apresentada na figura 3.8b, onde o teste efetuado em cada nó é ilustrado nos respetivos ramos.
A árvore de classificação é iniciada com uma raiz, onde são considerados todos os indivíduos da amostra de treino: 114 recém nascidos prematuros extremos, dos quais 61 morrem e 53 sobrevivem (61|53). Esta raiz é segmentada pela variável GA com um valor de corte de 25 weeks. Os 114 indivíduos são divididos em dois subconjuntos complementares caso tenham GA<25 weeks ou GA≥ 25 weeks. Para o caso em que GA≥25 weeks é necessário novamente decidir consoante a variável MJD Inborn Delivery e assim sucessivamente. No fim deste processo temos então uma regra combinada das variáveis presentes na árvore. Por exemplo, se um recém nascido prematuro extremo verificar GA≥25 weeks e MJD Inborn Delivery=Yes e Weight<680 g, é previsto como morto (**Dead**).

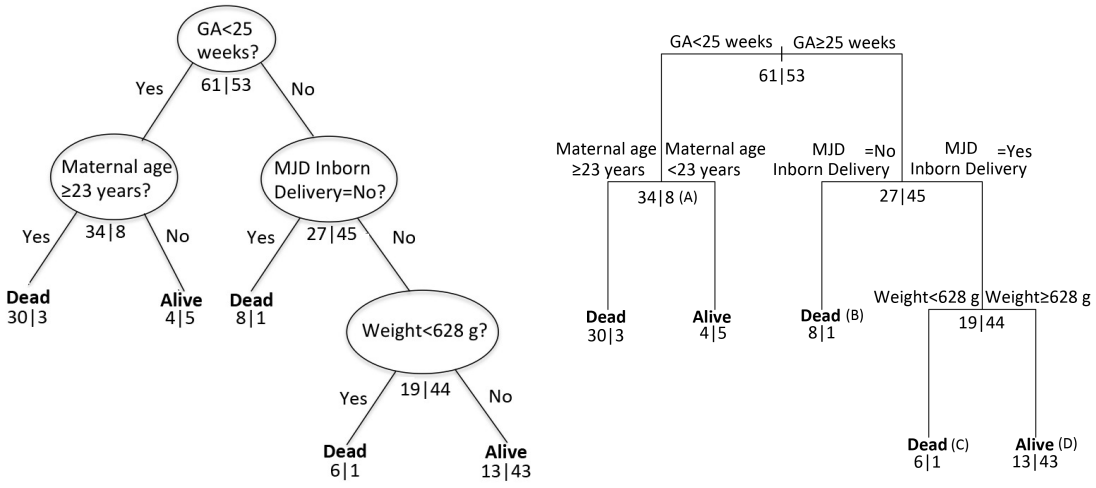



Figura 3.8: Duas possíveis representações para a árvore máxima de mortalidade. Onde, em cada nó se observa # **Dead** | # **Alive** na amostra de treino.

Após a compreensão da estrutura da árvore de classificação através do algoritmo de CART é indispensável perceber como são escolhidas as variáveis para segmentação de um nó bem como a estimação dos respetivos valores de corte.

**Segmentação de um nó**
A escolha da variável que deve segmentar um nó e a estimação do seu ponto de corte assume muita

importância na construção de uma árvore. Relembre-se que a construção baseada no algoritmo de CART apenas assume divisões binárias, pelo que só é necessário estimar um único ponto de corte em cada divisão.
Assuma-se $X = (X_1, ..., X_v, ..., X_p)$ como sendo o vetor das $p$ variáveis explicativas (contínuas e categóricas) presentes na amostra de treino. Consideremos a segmentação de um nó de acordo com a classificação da variável:

- **Variáveis categóricas nominais**
  Seja $X_v$ uma variável categórica, onde $B = \{b_1, b_2, ...b_L\}$ representa o conjunto das suas $L$ categorias. Para variáveis discretas, o método adotado sugere a divisão das categorias em dois subconjuntos complementares, associados respetivamente a cada um dos nós descendentes. Assim, consideram-se as possíveis $2^{L-1}-1$ divisões binárias associadas. Nestes casos os testes presentes em cada ramo da árvores serão da forma:

  $$\{X_v \in S\} \quad \text{e} \quad \{X_v \notin S\}$$

  em que $S$ é um dos subconjuntos possíveis de $B$. É de salientar que quando $L$ é elevado, o número de testes a considerar aumenta exponencialmente.

- **Variáveis contínuas ou categóricas ordinais**
  Seja $X_v$ uma variável contínua ou categórica ordinal contendo $N$ valores possíveis na amostra. Para este tipo de variáveis, o algoritmo de CART assume uma pesquisa exaustiva em vista a determinar os pontos de corte para tais variáveis. Note-se que são possíveis de obter $N$ segmentações diferentes, quando considerados os $N$ possíveis valores. Uma abordagem comum passa por ordenar os $N$ valores da variável $X_v$ $(X_{v,1}, X_{v,2}, ..., X_{v,N})$ e testar os atributos binários obtidos através do ponto médio entre 2 valores consecutivos:

  $$c_n = \frac{X_{v,n} + X_{v,n+1}}{2} \quad n = 1, ...N \tag{3.7}$$

  Assim, as possibilidades a serem consideradas são reduzidas para $N - 1$. O teste associado a este tipo de variável é da forma:

  $$\{X_v \leq c_n\} \quad \text{e} \quad \{X_v > c_n\}$$

Focando-nos no objetivo principal desta etapa, pretende-se então escolher a variável da amostra de treino mais relevante para segmentar um nó. Torna-se portanto essencial utilizar uma medida capaz de avaliar a qualidade da partição efetuada no nó por uma qualquer variável e, que permita a comparação de qualidade de partição por várias variáveis.
Breiman et al. (1984) sugeriram um *critério de impuridade*, estando este associado a uma medida de impureza calculada em cada nó. A ideia fundamental visa que considerando um nó $t$, os seus descendentes sejam mais *"puros"* do que o seu predecessor $(t)$. A ideia de mais *puros* refere-se a que os nós descendentes apresentem menos mistura de classes em comparação com o seu *nó pai*.
Geralmente, tal impureza é definida como:

- Definam-se as probabilidades de pertencer à classe $w_k$, $k = 1, 2, ...K$, num determinado nó $t$ como sendo $p_k = Prob(w_k|t)$, tais que:

  $$\sum_{k=1}^{K} p_k = 1 \tag{3.8}$$

- Designe-se por *i(t)* uma medida de impureza de um nó $t$, definida como uma função $\phi$ não negativa de $p_1, ..., p_K$, com as seguintes propriedades:

$$\phi(\frac{1}{K}, \frac{1}{K}, ..., \frac{1}{K}) = \quad \text{máximo} \tag{3.9}$$

$$\phi(1, 0, ..., 0) = \phi(0, 1, 0, ..., 0) = ... = \phi(0, 0, ..., 0, 1) = 0 \tag{3.10}$$

$$\phi \quad \text{é uma função simétrica em} \quad p_k, k = 1, ..., K \tag{3.11}$$

Verifica-se que a impureza num nó $t$ é máxima quando as $p_k$ são muito homogéneas nesse nó (equação (3.9)), indicando muita mistura de classes. Quando o nó contém todos os objetos centrados numa só classe, a impureza é nula e o nó diz-se *puro* (equação (3.10)).

As medidas de impureza mais comuns em árvores de classificação são:

$$\textbf{Índice de Gini} \qquad i(t) = 1 - \sum_{k=1}^{K} p_k^2 \tag{3.12}$$

$$\textbf{Entropia} \qquad i(t) = -\sum_{k=1}^{K} p_k \log(p_k) \tag{3.13}$$

No presente trabalho, as árvores de classificação serão obtidas pelo índice de Gini, por recomendação dos autores de CART.

Qual será então a variável a escolher para segmentar um determinado nó? Uma resposta óbvia será selecionar o atributo cuja partição em dois nós descendentes diminua o mais possível a impureza. O *decréscimo de impureza* é definido de acordo com o critério *i(t)*:

$$\triangle i(t) = i(t) - [p(t_l)i(t_l) + p(t_r)i(t_r)], \tag{3.14}$$

onde $t_l$ e $t_r$ representam os nós descendentes esquerdo e direito ao nó respetivamente, $i(.)$ as suas impurezas e $p(.)$ a proporção de indivíduos do nó $t$ que verificaram a decisão desse nó descendente (Duda et al., 2001). O algoritmo de CART irá então proceder a uma pesquisa exaustiva de todas as variáveis disponíveis assim como dos testes a elas associadas e escolherá para segmentar cada nó a variável que maximizar o critério (3.14).

Debrucemo-nos agora sobre a complexidade computacional de tal pesquisa. Para variáveis contínuas ou categóricas ordinais, o algoritmo reduz a sua pesquisa a $N - 1$ possibilidades de árvores binárias, uma vez que os $N$ valores tomados por estas variáveis são ordenados e se consideram os pontos médios de dois valores consecutivos. No entanto, quando estamos na presença de atributos categóricos nominais que assumem bastantes valores não ordenados a complexidade da pesquisa torna-se excessiva. Para contornar tal situação, é necessário ter em conta a seguinte proposição:

**Proposição 3.1** *(Breiman et al. (1984), Ripley (1996) pág.218)*
*Suponhamos que i(t) é uma função estritamente côncava.*

**(i)** *O decréscimo de impureza $\triangle i(t)$ é não negativo, e é zero sse as distribuições de probabilidade são as mesmas em todos os seus nós descendentes.*

**(ii)** *Suponhamos que K=2. Para uma variável categórica $X_v$ que assume L níveis, $\{b_1, b_2, ..., b_L\}$, ordene-se crescentemente as probabilidades $Prob(w_1|X_v = b_i)$:*

$$Prob(w_1|X_v = b_{i1}) \leq Prob(w_1|X_v = b_{i2}) \leq ... \leq Prob(w_1|X_v = b_{iL})$$

*Assim, uma divisão da forma* $\{b_{i1}, ..., b_{il}\}, \{b_{il+1}, ..., b_{iL}\}$*, maximiza o decréscimo de impureza.*

Assumir que a função de impureza $\phi$ é estritamente côncava é equivalente a dizer que $\phi'' < 0$ no domínio das probabilidades, tendo portanto um único valor máximo. A consequência do impacto desta proposição reside no facto de se verificar que qualquer que seja a partição efetuada, os nós descendentes serão sempre mais puros do que o nó pai (ponto **(i)**). A veracidade desta proposição permite ainda, para o caso de $K = 2$, reduzir o número de procuras sobre atributos categóricos de $2^{L-1} - 1$ para $L - 1$, passando-se de uma complexidade de procura exponencial para uma linear (ponto **(ii)**). Esta redução pressupõe a ordenação crescente das categorias de uma variável de acordo com a probabilidade destas numa das classes, por exemplo a primeira. Assim, as variáveis categóricas nominais conseguem ser tratadas de forma equivalente às variáveis contínuas, onde também existe uma ordenação subjacente. A prova da proposição 3.1 pode ser consultada em Ripley (1996).

As medidas de impureza, índice de Gini e Entropia (equação (3.12) e (3.13)) são funções estritamente côncavas e, portanto, a proposição 3.1 é válida. Contudo, outras funções de impureza poderiam ser utilizadas, como por exemplo o erro (custo) de classificação $i(t) = 1 - max(p_k)$. No entanto, $i(t)$ não é diferenciável e não valoriza tanto a presença de nós puros (Breiman et al. (1984)).

Considere-se a árvore obtida para um ano de mortalidade apresentada na figura 3.8b. Nesta situação estamos perante o caso em que $K = 2$ (**Dead ou Alive**).
Suponhamos que o nó $t$ que pretendemos segmentar diz respeito à raiz (61|53). O algoritmo de procura utilizado teve de realizar uma pesquisa para determinar qual a variável que deveria efetuar a divisão da raiz, considerando como função de impureza o índice de Gini. Se a título de exemplo compararmos a divisão escolhida com apenas uma outra possibilidade também testada pelo algoritmo, verificamos que de facto a variável GA com um ponto de corte de 25 weeks é ótima para a divisão do nó $t$. A figura 3.9 e a tabela 3.5 apresentam as partições a testar e os respetivos cálculos referentes aos decréscimos de impureza.
Observa-se que os decréscimos de impureza em ambos os casos (I e II) são pouco elevados (0.088 versus 0.073) (Tabela 3.5). Esta situação acontece frequentemente, uma vez que é difícil conseguir-se grandes decréscimos de impureza com apenas uma divisão. Embora $\triangle i(t)$ estejam muito próximos em ambos os casos, $\triangle i(t)$ é maior para a variável GA com um ponto de corte de 25 weeks e portanto, esta é a variável escolhida para a divisão do nó $t$ (raiz).

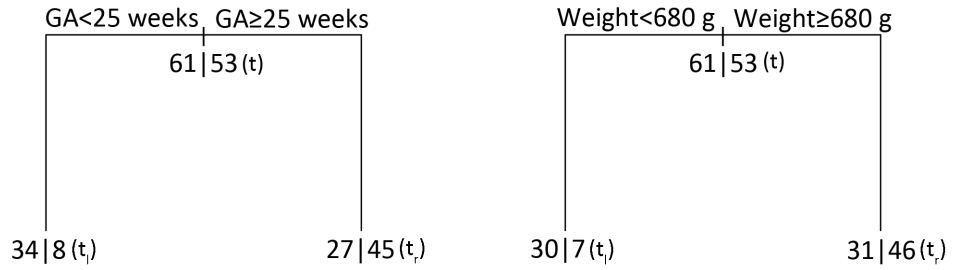



Figura 3.9: Dois exemplos de escolha de divisão de um nó através do índice de Gini. Caso I e II respetivamente.

Tabela 3.5: Cálculos do decréscimo de impureza efetuados referentes aos dois exemplos ilustrados na figura 3.9.

| **Caso I** | **Caso II** |
|---|---|
| $i(t) = 1 - \left[\left(\frac{61}{114}\right)^2 + \left(\frac{53}{114}\right)^2\right] = 0.498$ | $i(t) = 1 - \left[\left(\frac{61}{114}\right)^2 + \left(\frac{53}{114}\right)^2\right] = 0.498$ |
| $i(t_l) = 1 - \left[\left(\frac{34}{42}\right)^2 + \left(\frac{8}{42}\right)^2\right] = 0.308$ | $i(t_l) = 1 - \left[\left(\frac{30}{37}\right)^2 + \left(\frac{7}{30}\right)^2\right] = 0.307$ |
| $i(t_r) = 1 - \left[\left(\frac{27}{72}\right)^2 + \left(\frac{45}{72}\right)^2\right] = 0.469$ | $i(t_r) = 1 - \left[\left(\frac{31}{77}\right)^2 + \left(\frac{46}{77}\right)^2\right] = 0.481$ |
| $\triangle i(t) = 0.088$ | $\triangle i(t) = 0.073$ |

### 3.2.2 Critério de paragem

Uma decisão pertinente na construção das árvores refere-se à escolha do critério de paragem do procedimento recursivo do algoritmo de CART (Figura 3.7), tendo de se atribuir uma classe ao nó terminal. Várias hipóteses podem ser consideradas:

**(i)** Trivialmente, quando o nó é *puro*, centrando-se todas as observações numa dada classe.

**(ii)** Atribuir um limiar $\beta$ para o valor do decréscimo de impureza $\triangle i(t)$. Se o melhor atributo a segmentar o nó satisfaz $\triangle i(t) \leq \beta$, o crescimento para.

**(iii)** Atribuir um limiar $\gamma$ para o número de observações no nó. Se o número de observações num determinado nó for inferior a $\gamma$, o crescimento para.

A situação **(i)** é sempre considerada, já as hipóteses **(ii)** e **(iii)** são utilizadas consoante for necessário ou quando definido no problema em estudo.

**Atribuição de classe a um nó terminal**
Após a decisão de que um determinado nó $t$ é considerado como terminal, $t \in \tilde{T}$, é indispensável a atribuição de uma classe. As duas regras a admitir são as seguintes:

- **Atribuição da classe mais provável**
  Nesta situação, a classe a atribuir é a mais provável. Trata-se de uma regra simples que minimiza o erro de classificação.

- **Atribuição da classe baseada em custos**
  Esta regra visa ter um conta a matriz de custos associada ao problema em questão. Nesta abordagem, o intuito não será escolher a classe que tende a diminuir o erro de classificação mas sim atribuir a classe que minimizar o custo. Considerando um nó $t$ e tendo em conta as probabilidades *a posteriori* nele estimadas , $p_k$, o custo esperado de cada classe é dado pela seguinte expressão:

$$R(w_j|t) = \sum_{k=1}^{K} C(j|k) Prob(w_k|t), \quad k = 1, ..., K \tag{3.15}$$

  onde $C(j,k)$ é o custo de atribuir a classe $j$ quando a verdadeira é a $k$ e $Prob(w_k|t)$ é a probabilidade *a posteriori* de um indivíduo pertencer à classe $w_k$ no nó $t$. Intuitivamente,

atribuir-se-á a classe que minimizar a equação (3.15).

Destaque-se que o uso de custos de má classificação pode também ser utilizado durante a construção da árvore, tendo portanto influência no seu desenvolvimento. Nesta situação, as medidas de impurezas, são adaptadas à introdução de tal custo (Breiman et al., 1984; Ripley, 1996). Esta abordagem de custos durante a construção da árvore será introduzida na subsecção 3.2.5.

### 3.2.3 Poda

O procedimento para a construção de uma árvore de classificação permite obter uma árvore máxima $T_{max}$, com base na amostra de treino (Figura 3.7). Contudo, é notório que esta árvore poderá estar demasiado ajustada (*"overfitted"*) às características da amostra de treino, podendo não ter um desempenho adequado quando utilizada noutras amostras.
De forma a contornar esta situação, é necessária a utilização de um processo de poda. Este procedimento tem como objetivo a obtenção de uma árvore mais pequena do que a máxima pela eliminação de nós que demonstrem não ter grande relevância.

A árvore máxima pode ser podada através de dois métodos (Duda et al., 2001):

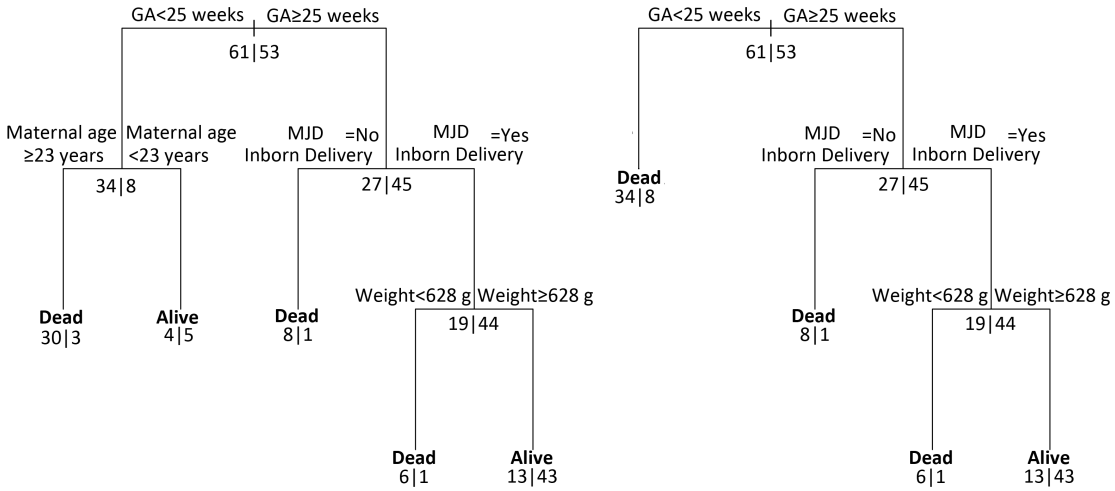
- **métodos pré-poda:** parar o crescimento da árvore por atribuição de um limiar $\beta$ para o decréscimo de impureza. Ressalve-se que esta abordagem já foi referida no ponto **(ii)** da subsecção 3.2.2.

- **métodos pós-poda:** deixar a árvore crescer o mais possível, $T_{max}$ e depois podá-la dando origem a uma subárvore.

Embora ambas as estratégias possam ser utilizadas, o método mais frequente é o pós-poda, sendo este adotado nos trabalhos de Breiman et al. (1984) e Quinlan (1993). A razão desta preferência advém da caraterística de que os métodos pré-poda param o crescimento da árvore quando não há uma divisão que decresça significativamente a impureza. No entanto, se dermos a possibilidade de estas divisões continuarem, posteriormente poderão surgir novas divisões com grandes decréscimos de impureza.

Na presente subsecção será abordada a técnica de poda proposta pelos autores de CART (Breiman et al., 1984): *poda por minimização do custo-complexidade*. Esta técnica é baseada no método pós-poda e consiste na eliminação de ramos não preditivos da árvore $T_{max}$.

A figura 3.10 ilustra o procedimento de poda considerando como exemplo a árvore obtida para descrever um ano de mortalidade. Da árvore máxima (Figura 3.10a) foi eliminado o nó segmentado pela variável Maternal age, dando origem à árvore podada (Figura 3.10b).




(a) Árvore máxima (b) Árvore podada

Figura 3.10: Processo de poda da árvore de classificação referente a um ano de mortalidade. O processo de poda identificou a variável Maternal age como a menos relevante na árvore de decisão.

Defina-se alguma notação necessária para a compreensão deste procedimento. Considere-se $\tilde{T}$ como sendo o conjunto dos nós terminais de uma árvore T. Seja $t$ um nó tal que $t \in \tilde{T}$.

Defina-se a estimativa do custo (erro) de classificação incorreta do nó $t$ por:

$$r(t) = \min_j \sum_{k=1}^{K} C(j|k)p(k|t), \quad k = 1, ..., K \tag{3.16}$$

Na presença de custos unitários, isto é $C(j|k) = 1$, $r(t)$ assume-se como:

$$r(t) = 1 - \max_k p(k|t), \quad k = 1, ..., K \tag{3.17}$$

Cada nó $t$ terá um impacto para o custo global da árvore de acordo com:

$$R(t) = r(t)p(t) \tag{3.18}$$

Generalizando, podemos definir o custo (erro) global de uma árvore $T$ como a soma do custo de todos os nós terminais:

$$R(T) = \sum_{t \in \tilde{T}} R(t) \tag{3.19}$$

A ideia fundamental da *poda por minimização do custo-complexidade* é a procura de árvores que permitam estabelecer um compromisso entre o custo de má classificação e a sua complexidade. Para qualquer subárvore de $T_{max}$, $T \preccurlyeq T_{max}$, defina-se:

- complexidade como sendo o número de nós terminais de $T$, ou seja, $|\tilde{T}|$;
- $\alpha \in [0, +\infty[$ como o parâmetro de complexidade;
- uma medida de custo-complexidade, $R_\alpha(T)$.

Formalmente,

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|, \tag{3.20}$$

onde $R(T)$ diz respeito ao custo de má classificação de $T$ na amostra de treino. Posteriormente, este processo consiste então em procurar para cada valor de $\alpha$ a subárvore $T(\alpha) \preccurlyeq T_{max}$ que minimize $R_\alpha(T)$. Assim, é obtida uma sequência de subárvores, donde posteriormente será escolhida a subárvore ótima.

Como $\alpha$ é um custo de complexidade associado a cada nó terminal, assim:

- se $\alpha = 0$, então a subárvore $T(\alpha) = T_{max}$;
- se $\alpha$ assumir valores pequenos, a penalização pela elevada complexidade da árvore é pequena e consequentemente $T(\alpha)$ será grande;
- se $\alpha$ assumir valores grandes, o fator complexidade será muito valorizado, fazendo com que a subárvore escolhida tenha poucos nós terminais.

Nestas circunstâncias, depreende-se que à medida que o valor de $\alpha$ aumenta, a complexidade da árvore (número de nós terminais) diminui.

O problema surge quando, para cada valor de $\alpha \in \mathbb{R}$ existe mais do que uma subárvore $T(\alpha)$ que minimize $R_\alpha(T)$. Vejamos como escolher tal subárvore.

**Definição 3.1** *(Breiman et al. (1984), pág.67)*
*A menor subárvore mínima (do inglês "smallest minimizing tree") $T(\alpha)$ para um determinado parâmetro de complexidade $\alpha$ é definida pelas seguintes condições:*

**(i)** $R_\alpha(T(\alpha)) = \min_{T \preccurlyeq T_{max}} R_\alpha(T)$

**(ii)** *Se $R_\alpha(T) = R_\alpha(T(\alpha))$, então $T(\alpha) \preccurlyeq T$.*

Pela definição 3.1 constata-se que se existir $T(\alpha)$, então ela é de facto única. Contudo, a proposição 3.2 afirma que de facto esta subárvore existe sempre.

**Proposição 3.2** *(Breiman et al. (1984), pág.68)*
*Para cada valor de $\alpha$ existe uma "smallest minimizing tree", como definido em 3.1.*

Está-se então em condições de perceber como se obtém a sequência de subárvores $T(\alpha)$ tendo em conta cada valor de $\alpha$. Esta sequência inicia-se considerando $\alpha = 0$. Pela equação (3.20) , tem-se que:

$$R_{\alpha=0}(T) = R(T_{max}). \tag{3.21}$$

O processo de obtenção da sequência de árvores inicia-se então com $T_{max}$ associada a $\alpha = 0$. Porém, Breiman et al. (1984) alertam que há limitações em começar a sequência com a árvore $T_{max}$, uma vez que poderá existir uma subárvore $T_1 = T(\alpha = 0)$ da árvore $T_{max}$ que tenha o mesmo erro de classificação incorreta, isto é, $R_\alpha(T_1) = R(T_{max})$.

Verifica-se que em qualquer árvore, para um nó $t$ não terminal:

$$R(t) \geq R(t_l) + R(t_r) \tag{3.22}$$

onde $t_l$ e $t_r$ representam os nós descendentes do nó $t$.

Para obtermos $T_1$ através de $T_{max}$, consideremos dois quaisquer nós terminais $t_l$ e $t_r$ de $T_{max}$ resultantes do nó $t$. Sempre que $R(t) = R(t_l) + R(t_r)$, os nós $t_l$ e $t_r$ são podados. Este processo é executado até não ser possível podar mais nós, e assim obtemos a subárvore $T_1$. Esta particularidade referida pelos autores de CART tem como propósito a eliminação de nós não informativos, permitindo que a sequência se inicie a partir de uma árvore mais pequena mas com o mesmo erro que $T_{max}$.

Assim, para qualquer nó $t \in T_1$, denotemos $T_t$ como sendo um ramo de $T_1$ com raiz em $t$. Consideremos ainda o nó $\{t\}$ que resultou da eliminação do ramo $T_t$, como sugere a figura 3.11. Neste exemplo, $t$ corresponde ao nó segmentado pela variável Maternal age, sendo $T_t$ o ramo que separa os prematuros consoante esta variável. Pretende-se determinar se este nó $t$ é ou não informativo na árvore de decisão.

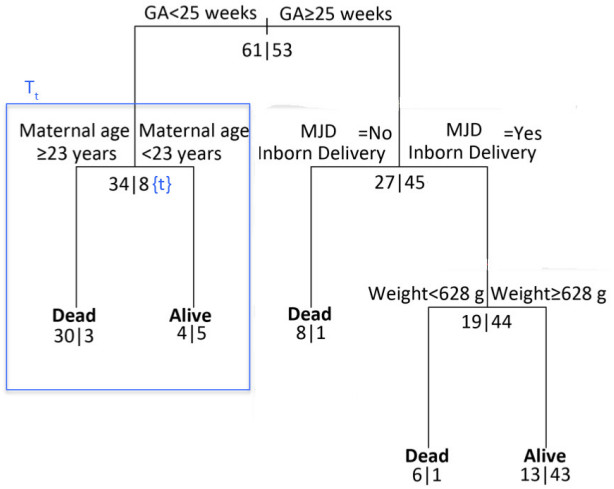



Figura 3.11: Árvore máxima $T_1$ destacando um nó candidato a ser podado.

A medida de custo−complexidade em $\{t\}$ e em $T_t$ é respetivamente:

$$R_\alpha(\{t\}) = R(t) + \alpha \quad \text{e} \quad R_\alpha(T_t) = R(T_t) + \alpha|\tilde{T}_t| \tag{3.23}$$

Caso

$$R_\alpha(\{t\}) \leq R_\alpha(T_t), \tag{3.24}$$

então compensa podar a árvore, pois o custo de classificação incorreta é menor considerando apenas $\{t\}$. É então crucial determinar qual o valor mínimo de $\alpha$ para o qual a poda é compensatória. Resolvendo a inequação (3.24) obtemos:

$$\alpha > \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1} \tag{3.25}$$

Assim, para obtermos uma subárvore $T(\alpha)$ basta numerarmos os nós das folhas para a raiz e caso a equação (3.24) seja satisfeita, podamos o nó $t$ em questão.
Nesta etapa, Breiman et al. (1984) sugerem a aplicação do algoritmo "elo mais fraco", que consiste em escolher para podar, o nó $t_1$ que está associado ao menor valor de $\alpha$, ou seja, $t_1$ é o primeiro nó para o qual $R_\alpha(\{t\}) = R_\alpha(T_t)$ [1]. Assim, obtemos a subárvore $T_2$ como sendo o resultado de $T_1 - T_{t_1}$,

[1] Se existir mais do que um nó que verique esta condição, podam-se todos ao mesmo tempo e não só $t_1$.

onde $T_{t_1}$ se refere ao ramo proveniente do nó $t_1$ e que foi podado por assumir um valor de $\alpha$ inferior. O novo parâmetro de complexidade associado a $T_2$ será então $\alpha_2$. Para obter $T_3$ e seguintes, existe um teorema em Breiman et al. (1984) e Ripley (1996) que garante não ser necessário recorrer à árvore $T_{max}$ para se obter a árvore seguinte, basta podar a árvore anterior que o resultado é o mesmo. Assim, este teorema garante que este procedimento é efetuado novamente mas tendo como ponto de partida $T_2$ e $\alpha_2$. O processo termina quando for encontrada uma subárvore que contém apenas um nó terminal. Este procedimento recursivo permite obter uma sequência de subárvores encaixadas, uma vez que em cada passo a árvore do passo anterior é podada. Os parâmetros de complexidade associados a estas árvores encaixadas também gozam de uma propriedade suportada pelo teorema 3.1, que refere o facto de quanto mais podada for a árvore (menor complexidade), maior será o seu parâmetro de complexidade.

**Teorema 3.1** *(Breiman et al. (1984), pág.71)*
*Os parâmetros de complexidade* $\alpha_k$, $k = 1, 2, ..., K$ *são uma sequência crescente, isto é,* $\alpha_k < \alpha_{k+1}$, $k \geq 1$, *onde* $\alpha_1 = 0$. *Para* $k \geq 1$, $\alpha_k \leq \alpha < \alpha_{k+1}, T(\alpha) = T(\alpha_k) = T_k$ [2]

Portanto, esta sequência de árvores e os respetivos parâmetros de complexidade são da forma:

$$T_{max} = T_1 \succ T_2 \succ T_3 \succ ... \succ T_K = \{t_1\} = \text{raiz} \tag{3.26}$$

$$T(\alpha_1) \succ T(\alpha_2) \succ T(\alpha_3) \succ ... \succ \text{raiz} \tag{3.27}$$

$$\alpha_1 < \alpha_2 < \alpha_3 < ... < \infty \tag{3.28}$$

**Escolha da subárvore ótima**
Após a obtenção da sequência de árvores encaixadas, é então necessário escolher de entre todas qual a melhor subárvore. Para tal, precisamos de uma estimativa "honesta"do custo de má classificação, $\widehat{R}(T_k)$, referente a cada árvore da sequência (Breiman et al., 1984), preferencialmente obtida num conjunto independente da amostra de treino, que foi utilizada para construir a árvore. Intuitivamente, escolher-se-á como classificador final a subárvore que minimizar tal estimativa de erro:

$$\widehat{R}(T_{k0}) = \min_k \widehat{R}(T_k), \tag{3.29}$$

onde $\widehat{R}(T_k)$ designa a estimativa do erro (custo de má classificação) da árvore $T_k$ e $T_{k0}$ representa a subárvore ótima.

Várias abordagens poderão ser utilizadas para estimar $\widehat{R}(T_k)$ (subsecção 3.1.1). Neste trabalho, por limitação do tamanho da amostra, optou-se por utilizar validação cruzada com $v$=10. Assim, considerando $\widehat{R}(T_k)$=$R^{CV}(T_k)$ e particularizando a equação (3.29) temos que:

$$R^{CV}(T_{k0}) = \min_k R^{CV}(T_k), \tag{3.30}$$

onde $R^{CV}(T_k)$ é uma estimativa do erro de validação cruzada da árvore $T_k$ que corresponde à média aritmética dos erros de $T_k$ em cada um dos $v = 10$ subconjuntos considerados.
Em suma, pretende-se escolher como subárvore ótima, aquela cujo respetivo erro seja minimizado. Contudo, esta escolha poderá muitas vezes ser "incerta", dado que as árvores são muito instáveis a pequenas mudanças no conjunto de dados. Neste sentido, os autores de CART sugerem a utilização da regra **1SE**, baseada nos erros padrão das estimativas. De acordo com Breiman et al. (1984), a subárvore ótima $T_{k1}$ será dada por:

[2] Note-se que neste contexto $K$ refere-se ao número de subárvores contidas na sequência, $k = 1, ..., K$ e não ao número de classes como tem vindo a ser utilizado nesta tese.

$$R^{CV}(T_{k1}) \leq R^{CV}(T_{k0}) + SE(R^{CV}(T_{k0})), \tag{3.31}$$

onde SE designa o erro padrão das estimativas dos erros de $T_{k0}$ nos $v = 10$ subconjuntos considerados para validação cruzada (do inglês, *"standard error"*). Assim, $T_{k1}$ é uma subárvore ótima se apresentar um erro de classificação não superior ao erro de classificação de $T_{k0}$ adicionado ao fator 1SE.

Vejamos um exemplo de determinação de subárvore ótima, com a árvore máxima de mortalidade (Figura 3.8), através do algoritmo de poda por minimização do custo-complexidade e validação cruzada com $v = 10$.
A figura 3.12 apresenta as estimativas de erro de validação cruzada de acordo com o número de nós terminais na árvore. É possível verificar que existe uma diminuição acentuada da média dos erros à medida que o tamanho da árvore aumenta. Seria de esperar que houvesse esta descida acentuada das estimativas de erro mas que em certo ponto estas ficassem mais estáveis e voltassem a aumentar gradualmente quando considerado um maior número de nós terminais (Breiman et al., 1984). Tal fenómemo não se verificou totalmente, talvez pelo facto de a árvore máxima não ser assim tão grande e consequentemente, possuir apenas 5 nós terminais. Adicionalmente, o facto de se ter efetuado validação cruzada pode-se tornar também um fator instável nas estimativas de erro. A utilização da regra **1SE** ajudará a reduzir esta instabilidade e a escolher como subárvore ótima uma árvore que tenha uma precisão semelhante à que assume ter um erro de validação cruzada inferior. Na figura 3.12 a linha a tracejado vermelha representa o valor de acordo com a regra **1SE** (termo do lado direito da equação (3.31)). Assim, a árvore podada a escolher encontra-se abaixo de tal limite delineado.
Analisemos a tabela 3.6, onde são apresentadas as estimativas de erros de validação cruzada bem como os respetivos erros padrão para cada uma das árvores presentes na sequência.

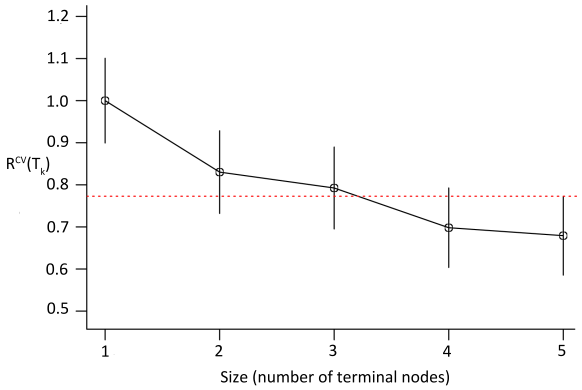



Figura 3.12: Estimativas do erro de validação cruzada em função do número de nós terminais (média± SE ).

Tabela 3.6: Estimativas de erros de validação cruzada para a escolha da subárvore ótima referente a um ano de mortalidade.

| $\alpha$ | size | $R^{CV}(T_k)$ | SE($R^{CV}(T_k)$) |
|---|---|---|---|
| 0.339623 | 1 | 1.00000 | 0.100479 |
| 0.132075 | 2 | 0.83019 | 0.098072 |
| 0.094340 | 3 | 0.79245 | 0.097177 |
| 0.018868 | 4 | 0.69811 | 0.094323 |
| 0.000000 | 5 | 0.67925 | 0.093642 |

Em conformidade com a regra **1SE**, a subárvore ótima é a menor árvore que apresentar um erro de validação cruzada inferior a 0.67925+0.093642=0.772892. Neste caso, corresponde à árvore com 4 nós terminais, à qual está associado $\alpha$=0.018868. A árvore podada obtida é a apresentada na figura 3.10b.

### 3.2.4 Valores em falta

A presença de valores em falta na amostra de treino e/ou na amostra de teste constitui um dos maiores problemas em modelos de classificação.
Uma das estratégias mais comum consiste em eliminar os valores em falta presentes na amostra, considerando para o estudo apenas aqueles que contenham informação completa em todas as variáveis. No entanto, esta opção pode desperdiçar bastante informação e só deverá ser considerada credível caso os valores desconhecidos tenham pouca representatividade (Duda et al., 2001).

Uma característica de destaque das árvores de classificação é a facilidade em lidar com a presença de tais valores (Ripley, 1996; Breiman et al., 1984; Duda et al., 2001). A não exclusão dos valores desconhecidos tem dois objetivos primordiais:

- utilização do máximo de informação possível na construção dos modelos;
- construção de uma árvore capaz de classificar novos objetos mesmo que estes apresentem valores desconhecidos em algumas variáveis.

Breiman et al. (1984) propuseram um algoritmo onde em cada nó são criadas divisões substitutas (*"surrogate splits"*) com desempenho semelhante à divisão original. Estas divisões alternativas, embora baseadas noutras variáveis têm uma associação preditiva semelhante no sentido em que os indivíduos que vão para o ramo esquerdo/direito são o mais possível os mesmos que com a divisão principal.

Quando ocorrer um valor em falta do atributo associado à divisão principal, será escolhido para este nó uma partição alternativa que melhor "imitar"a partição original, isto é, a que maximiza a probabilidade de ser realizada a mesma decisão do que a partição original nesse nó.

Para determinar tais partições substitutas é necessário definir uma medida de associação preditiva entre a divisão original e as respetivas partições substitutas (Breiman et al., 1984).

### 3.2.5 Modelos com custos de má classificação

Muitas vezes, o custo de má classificação de um indivíduo depende da sua classe. Por exemplo, prever um bebé como morto quando na verdade sobrevive, poderá traduzir um custo menor/maior do que classificar incorretamente um bebé como sobrevivente quando de facto o bebé morre.

A ideia da criação desta nova abordagem surgiu de uma reunião na MJD, onde os médicos foram questionados acerca do seu comportamento quando reportam o prognóstico do recém nascido prematuro aos pais. As opiniões foram um pouco divergentes. Alguns clínicos indicaram que devido ao prognóstico tão reservado que estas crianças apresentam, preferem evidenciar sempre o lado mais negativo quando comunicam aos pais, justificando que muitas vezes os pais criam expetativas demasiado "irreais"para a situação em que o filho se encontra. Por outro lado, existem clínicos que quando percecionam que a situação é mais agradável, optam por comunicar aos pais o lado mais positivo. No entanto, existem ainda casos em que os profissionais de saúde comunicam o prognóstico do recém nascido aos pais consoante o panorama lhes parecer mais positivo ou mais negativo.
Na tentativa de refletir estes comportamentos, torna-se equivalente a introduzir custos de má classificação, consoante o caso a reportar, sendo uma abordagem distinta da que tem vindo a ser feita até agora. Pretende-se reproduzir no método das árvores de classificação o que os médicos

fazem no seu dia a dia. Neste contexto, para a construção das árvores de classificação (mortalidade e morbilidade) foram criados dois modelos:

- **pessimista** que corresponde a sobrevalorizar o diagnóstico negativo;
- **otimista** que corresponde a sobrevalorizar o diagnóstico positivo.

Inicialmente, considere-se a seguinte matriz de custos e uma tabela de classificação:

$$C(j,k) = \begin{array}{c} \\ Dead/Severe \\ Alive/NonSevere \end{array} \begin{array}{c} \begin{array}{cc} Dead/Severe & Alive/NonSevere \end{array} \\ \begin{pmatrix} 0 & b \\ a & 0 \end{pmatrix} \end{array}$$

| | | Observed | | Overall |
|---|---|---|---|---|
| | | **Dead/Severe** | **Alive/Non Severe** | Accuracy(%) |
| Predicted | **Dead/Severe** | | | |
| | **Alive/Non Severe** | | | |
| Accuracy(%) | | | | |

Na matriz de custos, $C(j,k)$ representa o custo de atribuir (prever) a classe $j$ quando a verdadeira é a $k$. As linhas da matriz, $j$, correspondem às classes previstas que neste estudo são $j$=Dead, Alive (caso do estudo de mortalidade) ou $j$=Severe, Non Severe (caso do estudo de morbilidade). As colunas da matriz de custo, $k$, referem-se às classes verdadeiras (observadas), $k$=Dead, Alive (caso do estudo de mortalidade) ou $k$=Severe, Non Severe (caso do estudo de morbilidade). As entradas nulas da matriz de custo representam os casos em que as classes previstas coincidem com as observadas, por exemplo $C(Dead|Dead)$=0.
A atribuição dos custos será realizada tendo em vista o que se pretende observar na tabela de classificação.

No caso do modelo otimista, pretende-se aumentar as previsões corretas dos outcomes positivos (Alive/ Non Severe), isto é, aumentar a *Specificity* e no pessimista, aumentar as previsões corretas dos outcomes negativos (Dead/Severe), que equivale a aumentar a *Sensitivity*. Assim, com a introdução de custos, não estamos preocupados em que o erro de previsão global seja baixo, mas sim em aumentar as classificações corretas dos outcomes pretendidos. A ideia base será aumentar um custo relativamente a outro, sendo que a forma mais acessível é fixar um custo, por exemplo atribuindo o valor 1 e ir aumentando o outro. Neste trabalho, serão considerados apenas custos 2,3 e 4, por entendermos serem suficientes para determinar a diferença da gravidade na classificação. De forma a simplificar a explicação, reportamo-nos apenas ao caso da mortalidade.

Analisemos agora a situação otimista: temos como objetivo o aumento das previsões corretas na célula (Alive/Alive) da tabela de classificação. Para que tal aconteça, é preciso atribuir um custo maior a $C(Dead|Alive) = b$ de forma a que os recém nascidos prematuros sejam classificados na classe Alive. Neste contexto, fixamos $C(Alive|Dead) = a = 1$ e variamos $C(Dead|Alive) = b$, $b = 2, 3, 4$. Nesta abordagem otimista temos então que $C(Dead|Alive) = b \times C(Alive|Dead)$. É óbvio que o que pretendemos é que os recém nascidos nesta situação sejam todos classificados na célula (Alive/Alive) da tabela de classificação. Não obstante, ao atribuirmos um custo maior a atribuir a classe Dead quando a verdadeira é Alive, tem como consequência também aumentar a célula (Alive/Dead) da tabela de classificação.

A situação pessimista é exatamente o oposto da anterior: o intuito é aumentar as previsões corretas na célula (Dead/Dead) da tabela de classificação. Foi então fixado $C(Dead|Alive) = b = 1$ e variou-se o $C(Alive|Dead) = a$,$a = 2, 3, 4$.

Note-se que para a elaboração dos modelos otimistas e pessimistas os procedimentos foram semelhantes, apenas alteram os custos, consoante o estudo em causa. A variação dos custos teve como pretexto identificar qual deveria ser o mais apropriado em cada situação, uma vez que, por opinião médica não se conseguiu uma decisão assertiva. A implementação computacional deste estudo foi efetuada através do método de validação cruzada, *10-fold* e tendo em conta as seguintes etapas para cada custo estudado:

- foram obtidas as 10 árvores podadas.
- os valores do parâmetro de complexidade, $\alpha$ associados a cada uma delas.
- quando testadas nas respetivas amostras independentes, registaram-se o número médio de observações na linha de interesse da tabela de classificação (no caso do otimista é a linha Alive/Non Severe e no pessimista Dead/Severe) com o objetivo de se obter uma aproximação das classificações corretas e respetivos desvios padrão das 10 árvores.

Neste seguimento, para cada valor de custo atribuído foram calculados o $\alpha$, o número de observações na linha em questão e o desvio padrão através da média aritmética dos 10 valores obtidos para cada situação. O custo a ser escolhido será aquele a partir do qual se verificar um aumento da linha em questão da tabela de classificação.
Posteriormente, será contruída uma árvore de classificação baseada na amostra de treino, com base no custo escolhido e no parâmetro de complexidade associado. Os modelos serão avaliados através da tabela de classificação e de curvas ROC. No entanto, é importante frisar que as curvas ROC não são consideradas ideais para a avaliação de modelos com a presença de custos (Witten et al., 2011), pois o que se pretende estudar não é a precisão global mas sim o desempenho de classificação na classe em estudo. Neste contexto, as curvas ROC aqui apresentadas ilustram exatamente a tabela de classificação obtida em cada modelo. Os resultados obtidos por esta abordagem serão apresentados no fim da secção 3.2.6

### 3.2.6 Resultados

Nesta subsecção serão apresentados os resultados referentes à aplicação de árvores de classificação na previsão de:

- um ano de mortalidade (**Dead/Alive**)
- dois anos de morbilidade (**Severe/Non Severe**)

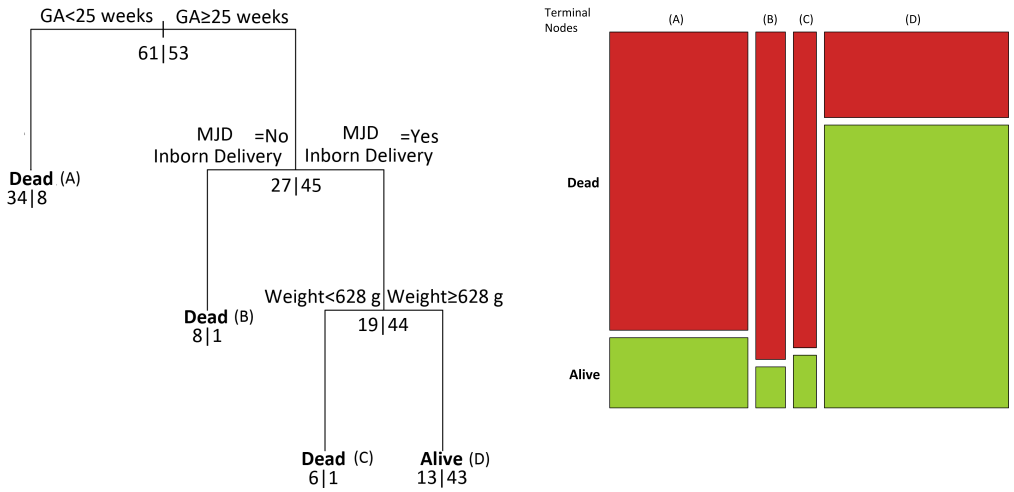
em recém nascidos prematuros extremos. As árvores de classificação para mortalidade e morbilidade serão avaliadas através da sua capacidade preditiva e com recurso a curvas ROC (Figura 3.7). A interpretação contextualizada do ponto de vista médico é apresentada posteriormente no capítulo 4.

Os modelos preditivos de mortalidade e morbilidade, foram obtidos considerando a seguinte conduta:

- amostras de treino e teste com valores em falta (subsecção 3.1.1);

- algoritmo de CART (secção 3.2) utilizando:
  - índice de Gini como medida de impureza (subsecção 3.2.1);
  - critério de paragem baseado num limiar de observações, $\gamma$, presente em cada nó (subsecção 3.2.2). Os cálculos encontram-se explicitados na tabela 3.1;
  - algoritmo de poda por minimização do custo-complexidade (subsecção 3.2.3);
  - obtenção da árvore podada ótima de acordo com a regra 1SE (subseção 3.2.3).

Parte do estudo de mortalidade foi utilizado como exemplificação de algumas etapas da parte teórica, pelo que se abdicará de repetir alguns passos para a obtenção da árvore.
A árvore máxima referente a um ano de mortalidade está representada na figura 3.10a, e destaca as variáveis GA, MJD Inborn Delivery, Weight e Maternal age como variáveis preditoras de mortalidade. Realce-se ainda que tais variáveis são recolhidas facilmente. Não obstante, esta árvore foi construida com a amostra de treino, podendo não ser a mais favorável de se utilizar para classificar novos recém nascidos prematuros extremos. Assim sendo, procedeu-se à poda da árvore, eliminando-se os nós não relevantes na previsão de mortalidade. A figura 3.13a apresenta a árvore podada, sendo possível constatar que a variável Maternal age foi excluída no processo de poda. De facto, esta exclusão não é surpreendente uma vez que por opinião clínica, tanto a variável como o ponto de corte de 23 anos na idade da mãe não é justificável.




(a) (b)

Figura 3.13: Árvore podada e mosaico referentes a um ano de mortalidade, onde em cada nó se observa # **Dead** | # **Alive** na amostra de treino.

A figura 3.13b apresenta o gráfico de mosaico que resume visualmente os resultados obtidos e representa a percentagem de indivíduos da amostra de treino classificados como **Alive** ou **Dead**, para cada nó terminal (A), (B), (C) e (D). Nesta situação confirma-se que o critério de atribuição de classe a um nó terminal é o da classe mais provável nesse mesmo nó.
Sublinhe-se que os três primeiros nós terminais, (A), (B) e (C) prevêm que os bebés sejam classificados como **Dead**. Já o último nó, (D) é o único que permite que o bebé seja classificado como **Alive**.

As variáveis importantes para mortalidade mostraram ser então GA, Weight e MJD Inborn Delivery, onde resumindo:

- Se GA<25 weeks → Dead
- Se GA≥ 25 weeks e MJD Inborn Delivery=No → Dead
- Se GA≥ 25 weeks e MJD Inborn Delivery=Yes e Weight< 628g → Dead
- Se GA≥ 25 weeks e MJD Inborn Delivery=Yes e Weight≥ 628g → Alive

Uma questão que merece destaque reside no facto de tanto a amostra de treino como a de teste possuirem valores em falta em algumas variáveis. Contudo, nenhum dos valores em falta corresponde às variáveis presentes na árvore de classificação. Desta forma, não foi necessário recorrer às partições alternativas para classificar os recém nascidos prematuros extremos da amostra de teste.

Obtida a árvore de classificação com base nas observações da amostra de treino, pretende-se agora avaliar a capacidade preditiva desta árvore quando aplicada às observações da amostra de teste. A previsão consiste então em verificar se a árvore construída é capaz de atribuir corretamente a classe dos novos indivíduos da amostra de teste. Os resultados obtidos são os que se reportam de seguida.

A figura 3.14 apresenta a tabela de classificação e a curva ROC da árvore de classificação de mortalidade. Constata-se que a precisão global deste modelo é de 69.4%, com 64.3% e 76.2% de previsões corretas nas classes Dead e Alive, respetivamente. Relativamente à curva ROC, verifica-se uma AUC de 70.2%), indicando que esta árvore de classificação tem um poder de discriminação aceitável (subsecção 3.1.4).

| | | Observed | | Overall |
|---|---|---|---|---|
| | | **Dead** | **Alive** | Accuracy(%) |
| Predicted | **Dead** | 18 | 5 | |
| | **Alive** | 10 | 16 | |
| Accuracy(%) | | 64.3 | 76.2 | 69.4 |

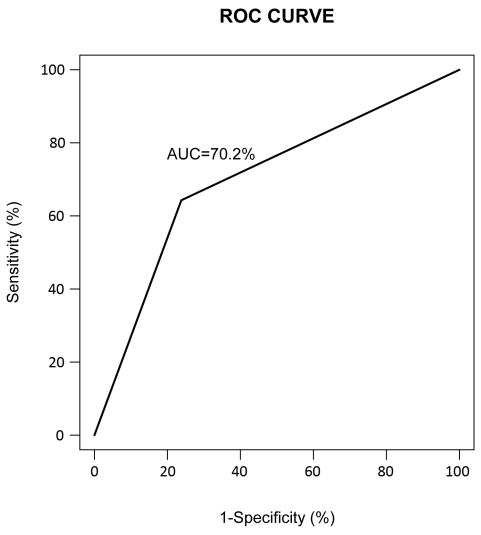



Figura 3.14: Tabela de classificação e curva ROC referentes a um ano de mortalidade, considerando árvores de classificação.

Os procedimentos para a construção da árvore de classificação para dois anos de morbilidade (Severe/Non Severe) foram análogos ao do estudo de mortalidade.

A figura 3.15 mostra a árvore estimada com base na amostra de treino, onde são identificadas as variáveis IVH, Weight, Maternal age, Caeserean Delivery e MR. É de realçar que a árvore apresenta nós puros, isto é, nestes nós a atribuição de classe não deixa qualquer dúvida:

- Se IVH=III or IV e Weight<885 g → Severe
- Se IVH=No or I or II e Maternal age ≥32 years e Caeserean Delivery=No → Non Severe

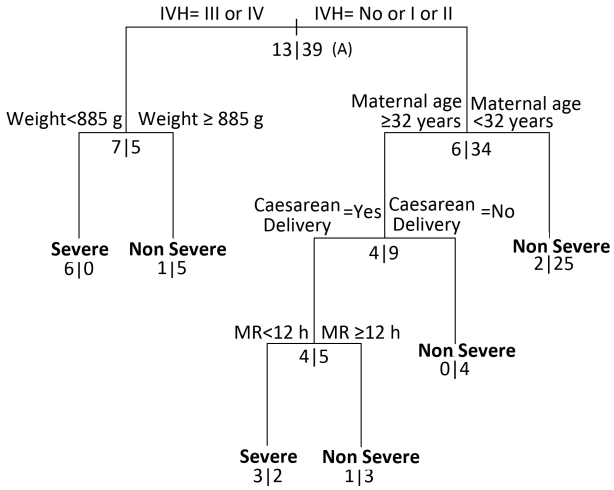



Figura 3.15: Árvore máxima referente a dois anos de morbilidade, onde em cada nó se observa # **Severe** | # **Non Severe** na amostra de treino.

De seguida, procedeu-se à poda da árvore. A figura 3.16 e a tabela 3.7 mostram as estimativas dos erros de validação cruzada.

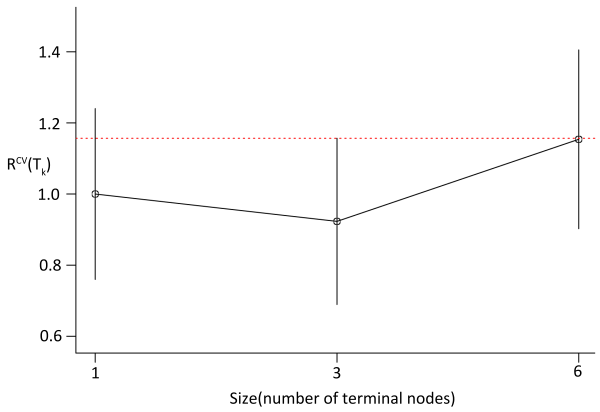


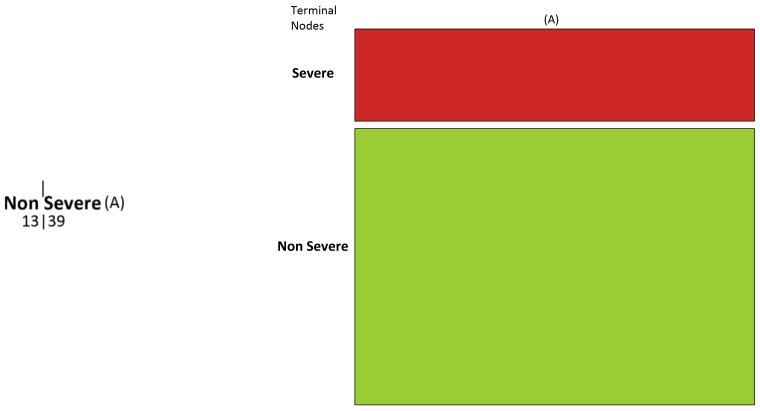

Figura 3.16: Estimativas do erro de validação cruzada em função do número de nós terminais (média±SE).

Tabela 3.7: Estimativas de erro de validação cruzada para a escolha da subárvore ótima referente a dois anos de morbilidade.

| $\alpha$ | size | $R^{CV}(T_k)$ | SE($R^{CV}(T_k)$) |
|---|---|---|---|
| 0.23077 | 1 | 1.00000 | 0.24019 |
| 0.025641 | 3 | 0.92308 | 0.23371 |
| 0.000000 | 6 | 1.15385 | 0.25131 |

Existem apenas três árvores possíveis: a árvore máxima com 6 nós terminais (Figura 3.15), uma árvore com três nós terminais e a que consta somente da raiz. Este resultado indica que o valor de $\alpha$ a partir do qual compensa podar demonstrou ser igual em vários nós, pois passamos imediatamente

de uma árvore com 6 nós para uma com 3. De acordo com a regra **1SE** (subsecção 3.2.3) a subárvore ótima corresponde a atribuir $\alpha = 0.23077$, estando este valor associado à árvore que contém apenas a raiz (Figura 3.17a). Assim sendo, nenhuma das variáveis que constavam na árvore máxima foram consideradas relevantes. Finalmente, todos os indivíduos serão classificados de acordo com a classe mais provável na raiz, **Non Severe** (Figura 3.17b).




(a) (b)

Figura 3.17: Árvore podada e mosaico referentes a dois anos de morbilidade, onde em cada nó se observa # **Severe** | # **Non Severe** na amostra de treino.

A figura 3.18 apresenta as métricas de avaliação do desempenho do modelo nulo na amostra de teste. A precisão global é de 90.9%, com 0% de classificações corretas na classe **Severe** e em contrapartida, 100% na classe **Non Severe**.

| | | Observed | | Overall |
|---|---|---|---|---|
| | | **Severe** | **Non Severe** | Accuracy(%) |
| Predicted | **Severe** | 0 | 0 | |
| | **Non Severe** | 2 | 20 | |
| Accuracy(%) | | 0 | 100 | 90.9 |

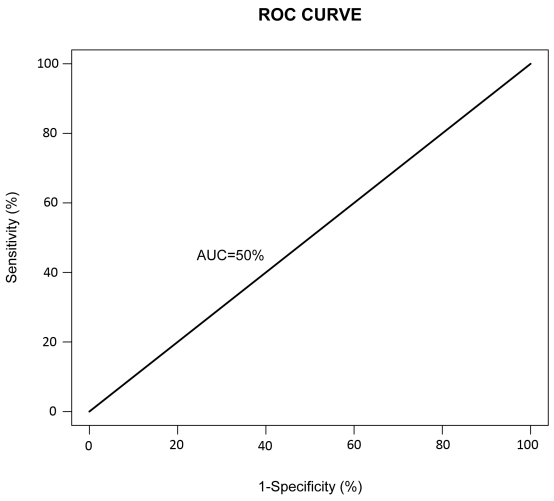



Figura 3.18: Tabela de classificação e curva ROC referentes a dois anos de morbilidade, considerando árvores de classificação.

No entanto, neste caso a previsão correta em 90.9% dos casos não traduz um modelo adequado de previsão. Como seria de esperar, a curva ROC traduz uma AUC de 50%.
O facto de não se ter identificado uma árvore de classificação de morbilidade, aliado ao facto de não se terem encontrado fatores de risco precoces (Capítulo 2, subsecção 2.4.2), indica que neste estudo não se encontrou evidência estatística de que alguma das variáveis tardiamente recolhidas estivesse associada a morbilidade. A curva ROC traduz uma AUC de 50%, indiciando não haver discriminação.

**Abordagem considerando custos de má classificação (resultados da subsecção 3.2.5)**
Foquemo-nos agora nos resultados desta análise para o caso da mortalidade. A tabela 3.8 apresenta para cada custo (primeira coluna), o número médio de indivíduos na linha das classificações corretas e o respetivo desvio padrão (segunda coluna). Adicionalmente, é também reportado o valor médio do custo de complexidade (terceira coluna).

Tabela 3.8: Custos a atribuir para obtenção do modelo otimista e pessimista referente a um ano de mortalidade.

Optmistic Model

| **Cost** | **Lines average (sd)** | **Cost-complexity average** |
|---|---|---|
| 1 | 6.5(1.958) | 0.051 |
| 2 | 8.3(2.003) | 0.098 |
| 3 | 10.5(2.461) | 0.195 |
| 4 | 10.7(2.214) | 0.182 |

Pessimistic Model

| **Cost** | **Lines average (sd)** | **Cost-complexity average** |
|---|---|---|
| 1 | 4.9(1.197) | 0.051 |
| 2 | 9.4(2.914) | 0.107 |
| 3 | 10.6(0.966) | 0.132 |
| 4 | 11(1.699) | 0.130 |

O custo 1 presente em ambas as tabelas (primeira linha) representa a abordagem com custos unitários, isto é, onde não se atribui custos específicos. Esta questão foi incluída no estudo com o objetivo de se apurar a partir de que custo a classificação correta de determinada classe é aumentada. Como já mencionado, os valores obtidos nas restantes colunas em ambas as tabelas, correspondem à média aritmética das 10 árvores construídas em cada custo. Foquemo-nos nas médias das linhas (segunda coluna): a amostra de treino para a mortalidade possui 114 indivíduos, pelo que quando se aplica validação cruzada, as amostras de teste terão cerca de 11 indivíduos. Neste sentido, o número máximo que se poderá obter será 11, daí serem apresentados valores próximos de 11 nessa coluna. De facto, a escolha do custo baseou-se no número médio de indivíduos na linha de interesse. No entanto, a ordem de grandeza do desvio padrão (sd) poderá não permitir aumentos significativos da média das linhas à medida que o custo aumenta. Estando cientes desta possível limitação, com a introdução de custos verifica-se que atribuir custo 2 em ambos os modelos otimista e pessimista é suficiente para se aumentar a classificação correta dos outcomes Dead/Alive, respetivamente. Esta conclusão retira-se, pois é notório que a média das linhas aumenta significativamente apenas com custo 2 (otimista: de 6.5 para 8.3 e pessimista de 4.9 para 9.4).

Para o modelo otimista foi contruída uma árvore de classificação em que se considerou $b = C(Dead|Alive) = 2$, $a = 1$, e parâmetro de complexidade, $\alpha = 0.098$. Já para o modelo pessimista, o parâmetro de complexidade foi de $\alpha = 0.107$, $a = C(Alive|Dead) = 2$ e $b = 1$. Como é evidente,

as árvores de classificação que se obtiveram diferem um pouco da árvore obtida para um ano de mortalidade (Figura 3.8), pois este método é bastante instável e a alteração de um pequeno parâmetro provoca imediatamente resultados distintos. Dado que a intenção deste estudo procura apenas determinar a partir de que custo teríamos as classificações corretas que pretendíamos, as árvores de classificação serão omitidas.

A Figura 3.19 apresenta as classificações obtidas para os modelos otimista e pessimista para a mortalidade e respetivas curvas ROC. Em comparação com a tabela de classificação referente a um ano de mortalidade sem introdução de custos (Figura 3.14), verifica-se que na abordagem otimista (cor azul) a classificação correta dos bebés da classe Alive aumentou cerca de 4.75% (de 76.2% para 80.95%) com a introdução de custo 2. No caso pessimista a classificação correta do outcome negativo é de 100%, indiciando que neste caso a árvore associada é uma raiz, não sendo este modelo significativo. Em ambos os modelos, a precisão global é bastante baixa (51.1% e 57.1%, respetivamente). As curvas ROC apresentadas sugerem que o modelo sem custos possui uma discriminação aceitável, o modelo otimista uma discriminação muito fraca e o pessimista não consegue discriminar nada.

| | | Observed | | Overall |
|---|---|---|---|---|
| | | **Dead** | **Alive** | Accuracy(%) |
| Predicted | **Dead** | **10**;**28** | **4**;**21** | |
| | **Alive** | **18**;**0** | **17**;**0** | |
| Accuracy(%) | | **35.7**;**100** | **80.95**;**0** | **51.1**;**57.1** |

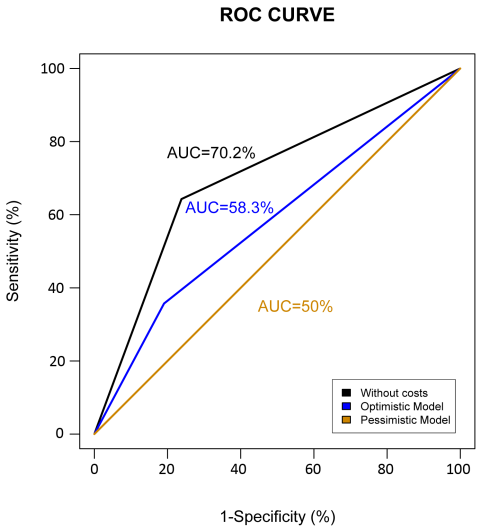



Figura 3.19: Tabela de classificação e curva ROC referentes a um ano de mortalidade, para o modelo otimista (azul) e pessimista (laranja).

As abordagens otimista e pessimista para o modelo de morbilidade foram efetuadas da mesma forma que na mortalidade. No entanto, apenas foi construído um modelo pessimista, uma vez que o modelo sem custos já previa todos os bebés na classe Non Severe (consultar Figura 3.18), e portanto já constituía o modelo mais otimista possível.

A tabela 3.9 apresenta os resultados alcançados para os custos considerados em estudo. Neste caso, a amostra de treino é constituída por 52 observações, pelo que cada amostra de teste em validação cruzada terá aproximadamente 5 indivíduos. Assim, o número máximo que poderá constar na média da linha de classificação de interesse é 5. Sem dar ênfase a nenhuma classe, isto é, considerando custos unitários, a média das linhas de prever na classe Severe é de 0.1. No caso do modelo pessimista, a penalização de custo 2, demonstra aumentar um pouco a classificação correta dos recém nascidos nesta classe. Esta situação de um aumento não muito significativo, já era de esperar dado que apenas 2 bebés integram esta classe. Deste modo, para o modelo pessimista a árvore foi construída atribuindo, $a = C(NonSevere|Severe) = 2$, $b = 1$ e um parâmetro de

complexidade de 0.254.

Tabela 3.9: Custos a atribuir para obtenção do modelo otimista e pessimista referente a dois anos de morbilidade.

Optimistic Model

| Cost | Lines average (sd) | Cost-complexity average |
|---|---|---|
| 1 | X | X |
| 2 | X | X |
| 3 | X | X |
| 4 | X | X |

Pessimistic Model

| Cost | Lines average (sd) | Cost-complexity average |
|---|---|---|
| 1 | 0.1(0.316) | 0.173 |
| 2 | 0.3(0.675) | 0.254 |
| 3 | 1.3(2.163) | 0.157 |
| 4 | 4.8(1.476) | 0.220 |

O desempenho da árvore de classificação construída nesta situação é ilustrado na figura 3.20. Confrontando os resultados agora alcançados, com os reportados na tabela de classificação apresentada na figura 3.18, constata-se que a percentagem de observações corretas aumentou em 50%, classificando corretamente apenas um recém nascido da classe Severe. O aumento torna-se significativo, uma vez que na amostra de teste de morbilidade apenas se tem 2 bebés nesta classe. Como consequência, a curva ROC associada a esta classificação demonstra ser muito discriminante (AUC=75%), quando comparada com a não discriminação do modelo sem introdução de custos.

| | | Observed | | Overall |
|---|---|---|---|---|
| | | **Severe** | **Non Severe** | Accuracy(%) |
| Predicted | **Severe** | **1** | **0** | |
| | **Non Severe** | **1** | **20** | |
| Accuracy(%) | | **50** | **100** | **95.5** |

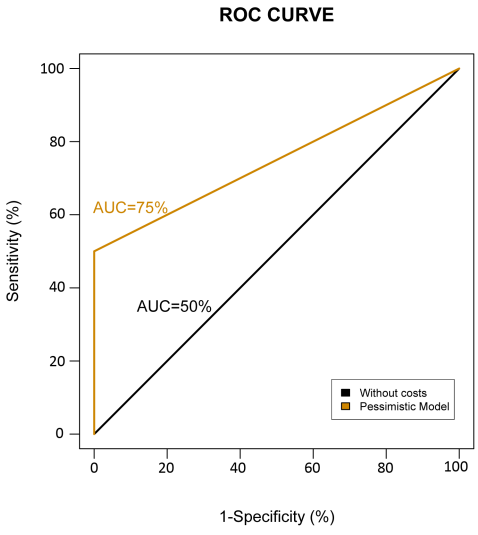



Figura 3.20: Tabela de classificação e curva ROC referentes a dois anos de morbilidade para o modelo pessimista.

## 3.3 Modelos baseados em regressão logística

Tal como na secção anterior, em que foram construídos modelos preditivos baseados em árvores de classificação, considerando as 26 variáveis da base de dados, é objetivo deste capítulo efetuar um estudo equivalente através da regressão logística.
Uma vez que este método já foi abordado com detalhe no capítulo anterior (Capítulo 2), apenas

se apresentará de forma sucinta a sua essência. Não obstante, a notação utilizada nos modelos de regressão logística com o intuito de previsão é bastante diferente da utilizada na deteção de fatores de risco. Assim, a notação deste capítulo, e em particular desta secção está de acordo com a utilizada em problemas de previsão, podendo ser consultada em Hastie et al. (2009)

A regressão logística, usada como modelo de previsão, consiste então em estimar diretamente as probabilidades *a posteriori*, $p_k$, associadas a cada valor da variável resposta ($K$ classes) através de funções lineares de $x$ e fazendo com que cada uma destas probabilidades corresponda a valores no intervalo [0,1]. Tais probabilidades são definidas como:

$$p_k = Pr(Y = w_k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \quad , k = 1, ..., K, \tag{3.32}$$

onde $x$ designa um objeto em estudo, $Y$ a variável correspondente às $w_k$ classes ($Y = \{w_1, ..., w_K\}$), $\pi_k$ a probabilidade à priori de pertencer à classe $k$ e $f_k$ a respetiva função densidade de probabilidade.

Recorde-se que nos casos em estudo temos apenas $K = 2$ , com classes representadas por $w_0$ e $w_1$ e portanto a probabilidade *a posteriori* que pretendemos obter é a que se refere ao evento em estudo (Dead ou Severe): $p_k = p_1$. Neste sentido, a equação de regressão é dada por:

$$logit(p_{1i}) = log\left(\frac{p_{1i}}{1 - p_{1i}}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ji} \quad j = 1, ..., p, \tag{3.33}$$

onde $i$ refere-se a um objeto, $\beta$ representa os coeficientes de regressão e $p$ as variáveis explicativas. Intuitivamente, efetuando cálculos na equação (3.33) , temos:

$$p_{1i} = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji}}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji}}} \tag{3.34}$$

### 3.3.1 Abordagem para procura exaustiva de modelos

A impossibilidade de aplicação de procedimentos do tipo *stepwise* aliados à necessidade de redução do número de variáveis, obrigou a explorar a associação entre variáveis, na tentativa de serem encontrados subconjuntos destas muito associados (subsecção 3.1.3). Porém, esta última hipótese também não foi conseguida. Assim sendo, é necessário adotar outra estratégia possível.

Considere-se um modelo de regressão logística tal como descrito na equação (3.33). De acordo com o descrito na secção 3.1.2, foi determinado para o caso da mortalidade e morbilidade, o número de variáveis a conter num modelo de regressão logística (Tabela 3.1). Todavia, este critério poderá ser relaxado, ou seja, poderão ser incluídas mais variáveis do que as determinadas. Esta situação motiva então procuras exaustivas de modelos multivariados.

Uma alternativa para se obter um modelo de regressão logística até $p$ variáveis será considerar o método de *seleção do melhor subconjunto de variáveis* (do inglês, *"Best Subset Selection"*) (Hosmer and Lemeshow, 2000; Hastie et al., 2009). O problema de seleção do melhor subconjunto passa por encontrar todos os $2^p$ subconjuntos possíveis de variáveis, e destes escolher o melhor de acordo com um determinado critério. Assim, esta abordagem efetua uma pesquisa exaustiva de modelos, determinando o melhor modelo para cada ordem[3] $j$, $j \in \{0, 1, 2, ..., p\}$. Intuitivamente, o modelo a

[3] A ordem do modelo refere-se ao número de variáveis nele contidas, por exemplo um modelo com duas variáveis é considerado um modelo de ordem dois.

escolher de entre todas as ordens será o que tiver um melhor desempenho de acordo com o critério adotado.

Um algoritmo eficiente e aplicado para a regressão linear é o designado *"branch and bound"* proposto por Furnival and Wilson (1974), podendo também ser consultado com detalhe no texto de Hastie et al. (2009). No entanto, Hosmer and Lemeshow (2000) citaram que este algoritmo pode também sofrer uma extensão para a regressão logística, caso sejam tidas em conta algumas etapas.

Na tentativa de encontrar tais subconjuntos, foi efetuada uma pesquisa acerca dos packages disponíveis no *Software R* para produzir tais resultados. Foram encontradas algumas rotinas que permitem selecionar subconjuntos de variáveis, sendo que a que nos pareceu mais adequada foi a rotina *bestglm()*. Esta função encontra-se disponível no *package* bestglm do *Software R*, permitindo para qualquer modelo linear generalizado selecionar o melhor subconjunto de variáveis para cada ordem e escolher de entre eles o melhor (McLeod and Xu, 2011).
Caso o modelo seja gaussiano (regressão linear) a rotina utiliza o algoritmo de *"branch and bound"*. Caso contrário, como na regressão logística, é aplicado um algoritmo proposto por Morgan and Tatar (1972), enumerando para todos os subconjuntos possíveis, $2^p$, as respetivas funções de log-verosimilhança.

Vários critérios podem ser escolhidos para a eleição do melhor subconjunto, sendo os mais frequentes: o critério de informação de Akaike (AIC) e o critério de informação Bayesiana (BIC) (Tabela 3.10). Ambos se baseiam na desviância, $D = -2\times$log-verosimilhança (Capítulo 2, subsecção 2.2.2) e no número de parâmetros[4] considerados no modelo, $k$.

Tabela 3.10: Expressões dos critérios de informação AIC e BIC.

| **AIC** | **BIC** |
|---|---|
| $D + 2k$ | $D + klog(n)$ |

Nesta investigação optou-se por escolher o critério BIC, uma vez que este penaliza também o tamanho da amostra de acordo com o número de parâmetros e a respetiva desviância. Portanto, admite uma penalização bastante superior à do AIC. Por esta razão, o critério BIC é reportado como sendo mais vantajoso, nunca selecionando modelos com mais parâmetros que o AIC (McLeod and Xu, 2011).

Resumindo, a procura dos subconjuntos de variáveis será efetuado considerando a rotina *bestglm()*:

- definiu-se que a ordem máxima do modelo seriam 15 variáveis, pois já é sabido que 26 são demasiadas;
- em todos os modelos de tamanho $j$, o parâmetro constante ($\beta_0$) será incluído;
- será efetuada uma pesquisa exaustiva de todos os subconjuntos de variáveis possíveis e calculada a respetiva função de log-verosimilhança;

[4]O número de parâmetros corresponde aos coeficientes de regressão estimados, sendo que variáveis politómicas assumem um parâmetro para cada dummy.

- para cada ordem, $j$, será escolhido o subconjunto que apresentar um critério BIC[5] menor;
- o melhor modelo, considerando todas as ordens, será o que apresentar um menor valor de BIC;
- será ainda possível ter-se noção dos 10 melhores modelos de todos os subconjuntos possíveis.

É aconselhável que este tipo de procura exaustiva seja efetuada quando o número de variáveis não é demasiado elevado (Guyon and Elisseeff, 2003; McLeod and Xu, 2011), pois o esforço computacional exercido é grande.

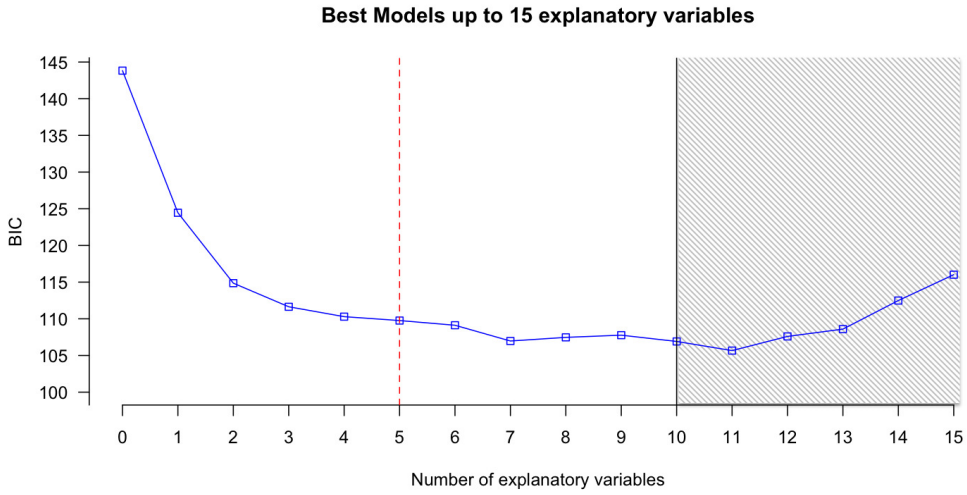
Para a obtenção de todos os resultados baseados neste método de seleção, foi necessário recorrer ao servidor do *CMUP*, para que a rotina *bestglm()* não necessitasse de ser interrompida. Este servidor apresenta uma enorme potencialidade. Particularizando as suas especificações, é um servidor com duplo processador de seis núcleos Intel Xeon X5650, com 48GB de memória RAM. Numa primeira experiência, considerou-se todas as procuras até 26 variáveis, chegando estes resultados a demorar cerca de 3 semanas a serem processados. Posteriormente, para além de se ter considerado apenas 15 variáveis foi dada prioridade a esta rotina no servidor, tendo demorado menos de uma semana. O tempo de execução foi sem dúvida uma dificuldade marcante neste trabalho.

### 3.3.2 Resultados

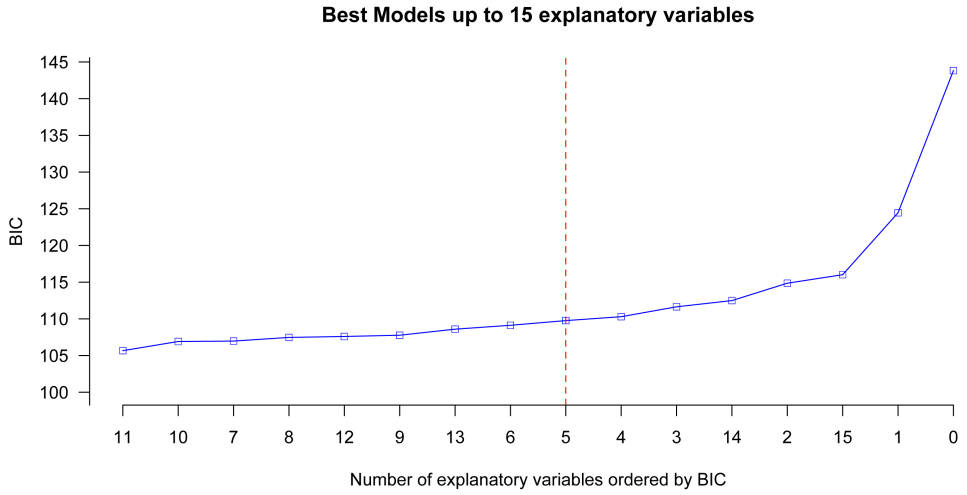
Estamos agora em condições de analisar os resultados para o modelo de um ano de mortalidade. Na figura 3.21 os resultados encontram-se explorados graficamente. A linha a tracejado vermelha indica o número máximo de variáveis que deveriam ser consideradas no modelo, de acordo com os cálculos efetuados na subsecção 3.1.2. A partir do modelo com 10 variáveis, surge o problema de se obterem modelos que não convergem, pelo que a análise foi restrita aos modelos que contêm entre 0 a 9 variáveis (Figura 3.21a). Na figura 3.21b, as ordens do modelo encontram-se por ordem crescente de valor de BIC, sendo que os melhores modelos são os primeiros a aparecer. Embora os valores de BIC não difiram muito nos primeiros modelos, os modelos com 10 e 11 variáveis revelam ser os melhores de acordo com o critério utilizado. No entanto, estes não serão escolhidos pelo facto de não convergirem. Assim, talvez a escolha fique restrita aos modelos precedentes a estes na figura 3.21b.

Os modelos considerando a inclusão de uma a nove variáveis encontram-se na tabela 3.12 (página 82), onde é apresentada para cada modelo a significância estatística das variáveis através do teste de Wald (Capítulo 2, subsecção 2.2.2).
Com a observação dos resultados, deparamo-nos que variáveis significativas como PDA, NIV, Infection, SGA apresentam coeficiente negativo, indicando que as crianças que possuem estas caraterísticas têm uma tendência maior para sobreviver do que os restantes. Este desfecho leva-nos a crer que à semelhança da BPD, estas variáveis não se encontram codificadas da melhor forma, para o caso da mortalidade. Sendo assim, o estudo destes modelos foi omitido, pela questão da codificação das variáveis não traduzir éticamente a realidade.

[5]O critério BIC sofre uma penalização relacionada com o número de parâmetros, $k$ (Tabela 3.10). É usual que a constante do modelo seja também considerada como um parâmetro. Todavia, a função *bestglm()* nunca a contabiliza, talvez pelo facto de esta ter sido imposta em todos os modelos no início da rotina.




(a)



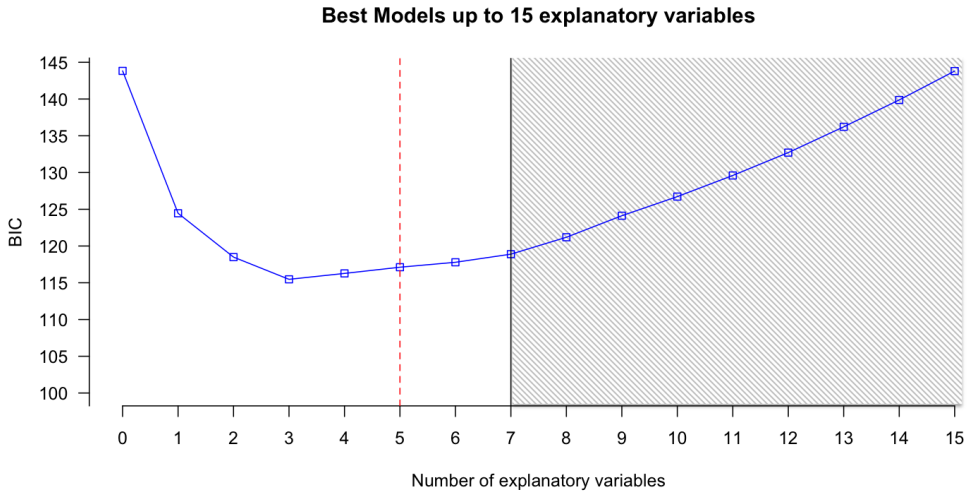

(b)

Figura 3.21: Valor do critério BIC para cada ordem do modelo de mortalidade: (a) de acordo com a ordem do modelo (b) ordenado por ordem crescente de valor de BIC.

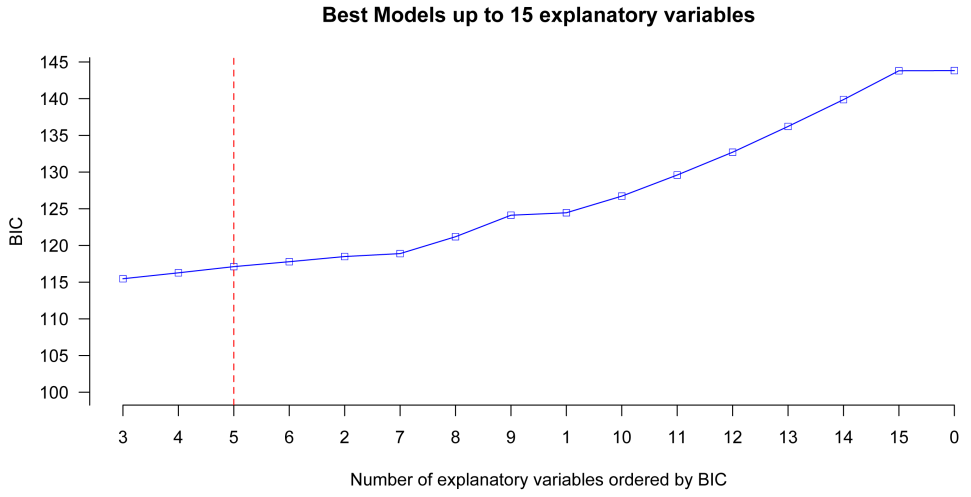
Para contornar esta situação, as variáveis mais promissoras de morbilidade (de acordo com opinião clínica e literatura na área média) foram retiradas da análise: NIV, HMD, IVH, LPV, Infection, PDA. Passamos assim a ter em estudo os mesmos 104 bebés (dimensão da amostra de treino de mortalidade, secção 3.1.1) mas apenas 19 variáveis.
Os resultados são discutidos seguidamente.

Pelo exposto na figura 3.22 conclui-se que de facto compensa ter um modelo, uma vez que o valor de BIC referente ao modelo nulo é superior a todos os restantes (Figura 3.22). Adicionalmente, constata-se que há uma tendência para que os valores de BIC voltem a aumentar, à medida que o número de variáveis a incluir no modelo também aumenta (Figura 3.22a). Esta situação pode ser justificada pela penalização que este critério faz ao número de parâmetros (Tabela 3.10).

Focando-nos especificamente no nosso objetivo, encontrar modelos até 5 variáveis, podemos observar que os melhores resultados são essencialmente obtidos em torno de 5, sendo os modelos com três e quatro variáveis considerados como os de eleição (Figura 3.22b). Todavia, os valores de BIC não são demasiado expressivos, sendo que se assemelham para os modelos com 3,4,5 e 6 variáveis. Neste estudo específico, conseguiu-se que todos os modelos até 15 variáveis convergissem, no entanto, a partir do modelo com 7 variáveis, inclusive, poucas ou nenhumas eram as variáveis neles significativas. Desta forma, só foram considerados para análise modelos com ordens inferiores a 7 (Figura 3.22a).




(a)




(b)

Figura 3.22: Valor do critério BIC para cada ordem do modelo de mortalidade, após exclusão das variáveis mais promissoras de morbilidade: (a) de acordo com a ordem do modelo (b) ordenado por ordem crescente de valor de BIC.

Embora o modelo com 3 variáveis tenha sido considerado o eleito, pelo método de seleção do melhor

subconjunto de acordo com o critério adotado (Figura 3.22b), é necessário ter algum cuidado com a análise dos restantes modelos. Interessa-nos que o modelo traduza a realidade mas também que seja constituído por variáveis que tenham alguma relevância clínica, podendo muitas vezes, este ser o argumento de seleção entre os melhores modelos.

A tabela 3.13 (página 83) apresenta os sete modelos propostos para estudo. As variáveis peso e idade gestacional mostraram ter significância inferior a 0.01 na maioria dos modelos e sugeriram que à medida que a resolução destas variáveis aumenta uma unidade, o risco dos recém nascidos prematuros morrerem é menor (ambas apresentam coeficientes negativos). Também o local de parto (MJD Inborn Delivery) foi uma variável sistematicamente identificada nos modelos, apresentando-se sempre como significativa.
Requerendo especial atenção aos dois modelos que apresentaram menor valor de BIC, modelos com 3 e 4 variáveis (Figura 3.22b), verifica-se que a diferença entre estes é apenas na variável ETT Resuscitation, que por sua vez é não significativa no modelo de ordem 4. A não significância desta variável, e a extrema importância que as restantes 3 variáveis oferecem, é mais que motivo para que o modelo de ordem 3 seja sem dúvida o eleito para prever mortalidade ao um ano. A equação deste modelo é então:

$$logit(p_1) = 27.47 - 5.65 \times 10^{-3} \times \text{Weight} - 0.82 \times \text{GA} - 2.78 \times \text{MJD Inborn Delivery} \quad (3.35)$$

Após a escolha do modelo "ideal"para caraterizar um ano de mortalidade, pretende-se analisar os dez melhores modelos representantes de toda a pesquisa exaustiva. O intuito desta análise será avaliar a credibilidade do modelo acima escolhido, tendo em conta os melhores modelos que poderiam ser utilizados.

Na tabela 3.14 (página 83) encontram-se explicitados os resultados obtidos, sendo que os modelos se encontram ordenados segundo o critério BIC.
Constata-se que os dez melhores modelos considerados na globalidade da pesquisa exaustiva possuem, na sua maioria, um número de variáveis igual ou inferior a 5 (valor recomendado).
Adicionalmente, verifica-se que as variáveis Weight, GA e MJD Inborn Delivery são constituintes de quase todos os modelos, apresentando-se sempre como significativas. Estes apenas diferem, no acréscimo de uma ou duas variáveis, para além das três anteriores. Note-se também que, como era esperado, os dois melhores modelos apresentados, dizem respeito aos melhores modelos de ordem 3 e 4, e, por isso são os que contam na figura 3.22 e na tabela 3.13.

A elevada importância que as variáveis Weight, GA e MJD Inborn Delivery revelaram ter, indicam que o modelo da equação (3.35) será um bom candidato para prever um ano de mortalidade.
Este modelo mostrou fazer um bom ajustamento aos dados quando acedido pela estatística de Hosmer and Lemeshow ($p - value = 0.1697$). As estatísticas de Cox&Snell $R^2$ e Nagalkerke's $R^2$ apresentaram valores de 33.4% e 44.6%, respetivamente.

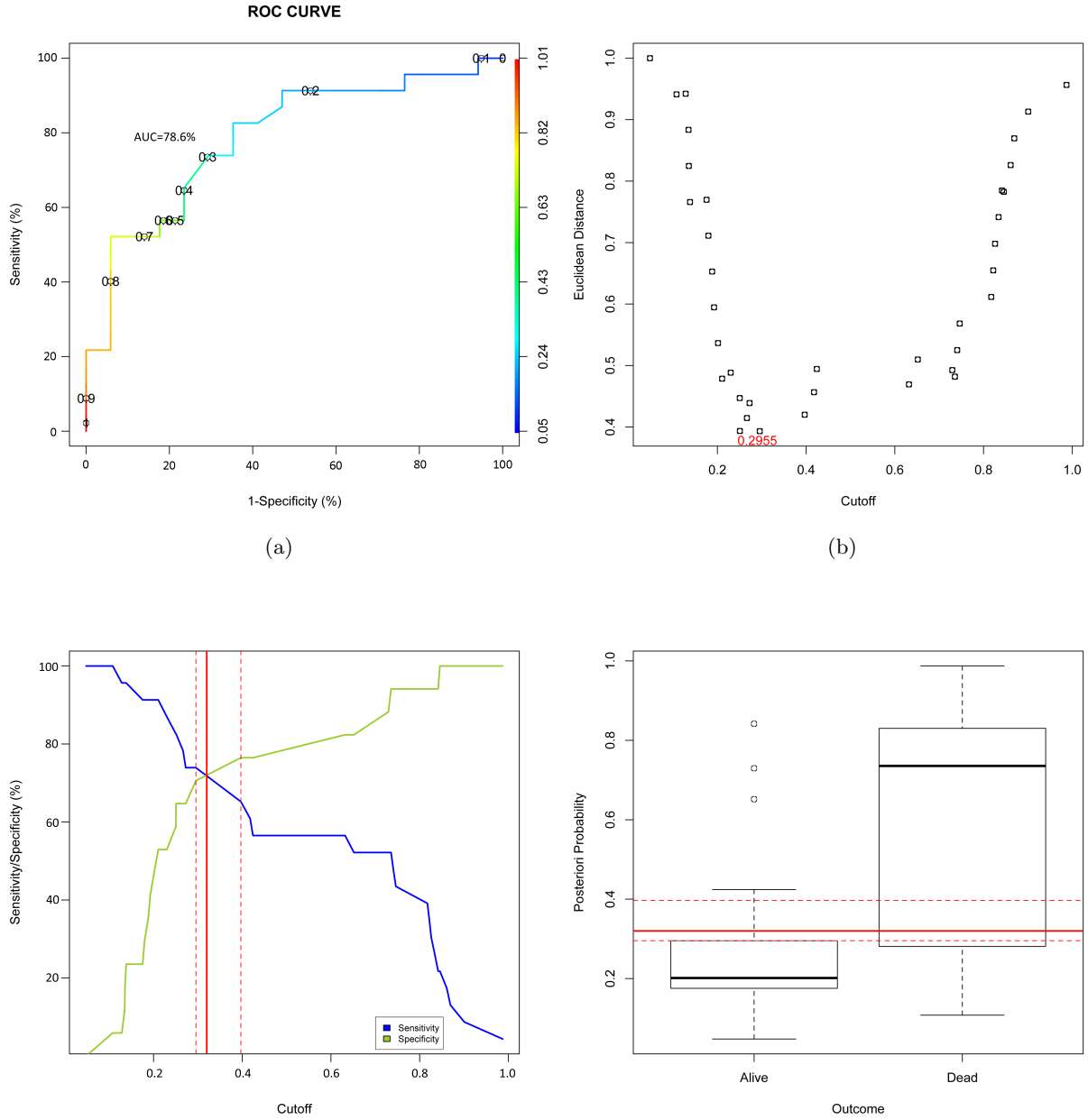
Finda a avaliação do ajustamento e da qualidade do modelo, é nosso propósito perceber a capacidade preditiva deste, na previsão de um ano de mortalidade em novos recém nascidos prematuros (neste caso, constituintes da amostra de teste).

A curva ROC sugere uma AUC=78.6% (Figura 3.23a), indicando que é possível ser feita uma discriminação aceitável entre as duas classes : Dead e Alive (consultar subsecção 3.1.4).

A vantagem poderosa das curvas ROC para classificadores que retornem como resultado de previsão a probabilidade *a posteriori*, $p_k$, é sem dúvida, poder-se avaliar as classificações corretas do modelo de acordo com o ponto de corte (*cutoff*) considerado. Cada discretização visível na curva ROC (Figura 3.23a), está associada a um ponto de corte específico, que por sua vez corresponde às probabilidades *a posteriori* estimadas.
Adicionalmente, esta figura apresenta uma escala colorida, onde as nuances de cores se referem a intervalos de pontos de corte, estando alguns destes realçados ao longo da curva. Por exemplo, para pontos de corte muito elevados, utilizam-se as cores alaranjadas. Nesta situação, a *Sensitivity* e *1-Specificity* tomam valores bastantes baixos.

A procura de um ponto de corte considerado ótimo, pode ser baseada em diversos critérios, consoante o interesse do estudo.
Neste trabalho, é nosso propósito determinar o ponto de corte que maximize a Sensitivity e por sua vez, minimize 1-Specificity. Assim, o ponto de corte que pretendemos encontrar e que consegue retratar este objetivo, será o que se situar mais próximo do ponto (0,100) (Figura 3.23a). Este critério de escolha do ponto de corte ótimo é mencionado usualmente em artigos relacionados com curvas ROC (Fawcett, 2006; Prati et al., 2008).
Dada a impossibilidade de este ponto ser determinado instantaneamente pela observação da figura 3.23a, optou-se por determiná-lo analiticamente. Para cada ponto de corte obtido, calculou-se a distância euclideana ao ponto (0,100) (Figura 3.23b), concluindo-se que segundo o critério adotado, o ponto de corte de 0.2955 é o valor na discretização que é ótimo.

Todavia, o ponto de corte ótimo poderá não estar restrito a ser um valor observado, até porque as discretizações, por vezes, conduzem a intervalos de *cutoff* bastante distantes. Relembrando que o que se pretende será encontrar o ponto de corte ótimo que maximize a *Sensitivity* e minimize *1-Specificity*, torna-se equivalente a procurar o *cutoff* que maximize a *Specificity* e a *Specificity* ao mesmo tempo. Esta questão, motivou então o estudo da interseção destas duas curvas (Figura 3.23c), onde o ponto de corte considerado ótimo será aquele onde as duas curvas se intersetam (Hosmer and Lemeshow, 2000).
A estimativa do ponto de corte foi calculada numericamente, considerando-se a interseção das retas que passam nos pontos de **Sensitivity/Specificity** dos pontos de corte 0.29 e 0.39 (linhas a tracejado, Figura 3.23c). Da resolução analítica, constatou-se que tal ponto de corte corresponde a 0.32 (linha vermelha, Figura 3.23c). No entanto, é necessário estar ciente de que se trata apenas de uma estimativa, havendo outros *cutoffs* possíveis, nomeadamente entre as linhas que se encontram a tracejado. Alternativamente, também se poderia ter optado por considerar como ponto de corte ótimo o ponto médio entre as bandas dos pontos de corte 0.29 e 0.39.
Por fim, após a decisão de qual deveria ser o ponto de corte ótimo, foram comparadas as probabilidades *a posteriori* estimadas ($\Leftrightarrow$ pontos de corte) com a verdadeira classe de cada indivíduo (Figura 3.23d). O boxplot indica que os prematuros extremos do evento que pretendemos prever (mortalidade) apresentam probabilidades bastante superiores aos restantes. As linhas vermelhas a tracejado são as correspondentes à banda de possíveis *cutoffs* e a linha a vermelho corresponde ao valor de 0.32, indicando que este *cutoff* parece conseguir distinguir a classe da maioria dos indivíduos, à exceção de alguns outliers.




(a) (b)

(c) (d)

Figura 3.23: Curvas ROC para o estudo de um ponto de corte ótimo referente ao modelo de morbilidade.

Outra solução alternativa para a identificação do ponto de corte ótimo poderia ser adotar aquele cujo desempenho global fosse maximizado (Figura, 3.24). Curiosamente, o ponto de corte 0.32 corresponde a uma estimativa de valor, para o qual a precisão consegue atingir valores mais elevados. Note-se que as bandas correspondem novamente aos valores de corte obtidos (0.29 e 0.39), e a linha a vermelho ao valor 0.32.

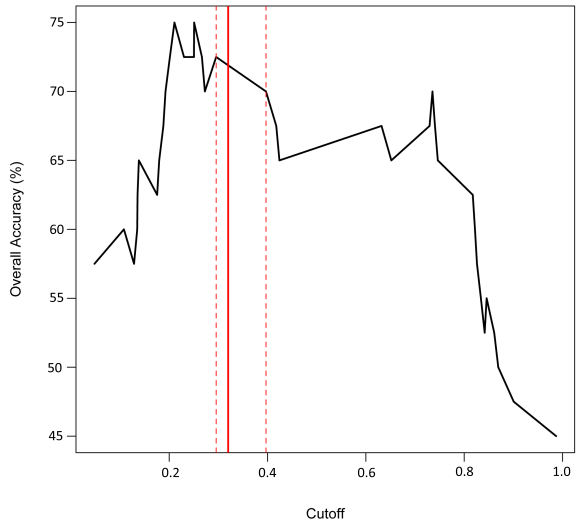



Figura 3.24: Precisão global do modelo da equação (3.35) de acordo com o ponto de corte.

Com a determinação do ponto de corte de 0.32, conseguiu-se obter uma classificação global correta de 70.0%, correspondendo a 65.2% de *Sensitivity* e 76.5% de *Specificity* (Tabela 3.11). Este resultado pode ser consultado na figura 3.23a, onde o ponto de corte 0.32 pertence à nuance esverdeada. Note-se que a precisão que se obteve não é exatamente a intersetada no gráfico anterior (Figura 3.24), pois entre os pontos de corte obtidos (0.29 e 0.39) não existe mais nenhuma discretização, sendo a precisão apresentada resultante da união das precisões entre os pontos.

Tabela 3.11: Tabela de classificação referente a um ano de mortalidade para um ponto de corte de 0.32, considerando o método da regressão logística.

| | | Observed | | Overall |
|---|---|---|---|---|
| | | **Dead** | **Alive** | Accuracy(%) |
| Predicted | **Dead** | 15 | 4 | |
| | **Alive** | 8 | 13 | |
| Accuracy(%) | | 65.2 | 76.5 | 70.0 |

No entanto, é sempre vantajoso aceder-se à performance do modelo, adotando um ponto de corte que seja comum, para que este modelo seja comparável com outros (Fawcett, 2006). O mais usual é considerar-se para tal, o valor de 0.5. Com este limiar, o modelo de mortalidade apresentou uma percentagem global de classificações corretas de 67.5%. A *Sensitivity* apresentou ser mais baixa com um valor de 56.5% e, consequentemente, a *Specificity* aumentou para 82.4% ($\Leftrightarrow$ *1-Specificity*=17.6%). Esta situação é identificada na figura 3.23a e torna-se equivalente a afirmar que com este ponto de corte, o modelo prevê melhor a classe Alive do que a Dead.

A título de curiosidade, caso se pretende-se efetuar um estudo equivalente ao realizado nas árvores de decisão, criando modelos otimistas e pessimistas, apenas se teria de alterar os pontos de corte para conseguir o pretendido. Por exemplo, para otimista, em que o intuito visa aumentar as classificações corretas da classe Alive, quanto mais alto fosse o ponto de corte melhor se traduziriam tais classificações. O procedimento inverso resultaria num modelo pessimista.

Tabela 3.12: Modelos de mortalidade e respetiva significância de variáveis de acordo com o teste de Wald.

| Model order | Weight ($\times 10^{-3}$) | Surfactant | PDA | GA | MJD Inborn Delivery | O2 | Infection | NIV | IVH=I or II | IVH=III or IV |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -7.37** | – – – | – – – | – – – | – – – | – – – | – – – | – – – | – – – | – – – |
| 2 | -8.33** | – – – | – – – | – – – | – – – | – – – | -1.81** | – – – | – – – | – – – |
| 3 | -8.68** | – – – | – – – | – – – | -2.61* | – – – | -1.72* | – – – | – – – | – – – |
| 4 | -6.72** | 2.74* | -2.12** | -1.02** | – – – | – – – | – – – | – – – | – – – | – – – |
| 5 | -7.02** | 2.47* | -1.93** | -1.00** | -2.24 | – – – | – – – | – – – | – – – | – – – |
| 6 | -14.17** | 2.88* | -2.00** | – – – | – – – | – – – | -1.69* | – – – | -0.99 | 2.14* |
| 7 | -10.28** | 3.99* | -2.78** | -1.54** | -3.82** | 5.08* | – – – | – – – | – – – | – – – |
| 8 | -10.33** | 4.04* | -2.86** | -1.61** | -3.76** | 5.36** | – – – | -2.35 | – – – | – – – |
| 9 | -7.92* | 4.41* | -3.71** | -2.04** | – – – | 5.78** | – – – | -22.8 | -2.05 | 2.01* |

**p-value<0.01, *p-value<0.05; Reference class for IVH is non IVH, for Epoch is (2000-2002), for MJD Inborn Delivery is other

| Model (cont.) order | Epoch (2003-2006) | Epoch (2007-2009) | Primipara | Maternal Age | SGA | Constant |
|---|---|---|---|---|---|---|
| 1 | – – – | – – – | – – – | – – – | – – – | 5.71** |
| 2 | – – – | – – – | – – – | – – – | – – – | 7.52** |
| 3 | – – – | – – – | – – – | – – – | – – – | 10.10** |
| 4 | – – – | – – – | – – – | – – – | – – – | 29.19** |
| 5 | – – – | – – – | – – – | – – – | – – – | 31.24** |
| 6 | – – – | – – – | – – – | – – – | -2.33* | 10.21** |
| 7 | – – – | – – – | -2.01* | – – – | – – – | 43.56** |
| 8 | – – – | – – – | -2.16* | – – – | – – – | 45.36** |
| 9 | 0.48 | 20.77 | – – – | 0.19* | – – – | 43.25** |

institutions and for binary variables yes/no is no

Tabela 3.13: Modelos de mortalidade após exclusão das variáveis mais promissoras de morbilidade e respetiva significância de variáveis de acordo com o teste de Wald.

| Model order | Weight ($\times 10^{-3}$) | GA | Maternal Age | ETT Resuscitation | Surfactant | O2 | MJD Inborn Delivery | Constant |
|---|---|---|---|---|---|---|---|---|
| 1 | -7.37** | – – – | – – – | – – – | – – – | – – – | – – – | 5.71** |
| 2 | -7.83** | – – – | – – – | – – – | – – – | – – – | -2.74* | 8.54** |
| 3 | -5.65** | -0.82** | – – – | – – – | – – – | – – – | -2.78* | 27.47** |
| 4 | -5.59** | -0.74* | – – – | 1.12 | – – – | – – – | -2.88* | 24.80** |
| 5 | -5.24** | -0.87** | – – – | 1.17* | – – – | 2.67 | -3.23** | 25.44** |
| 6 | -5.46* | -1.03** | 0.10 | 1.53* | – – – | 3.27 | -3.73** | 26.27** |
| 7 | -5.42* | -1.17** | 0.13* | 1.35* | 2.35 | 3.66 | -3.72** | 26.79** |

**p-value<0.01, *p-value<0.05;Reference class for MJD Inborn Delivery is other institutions and for binary variables yes/no is no

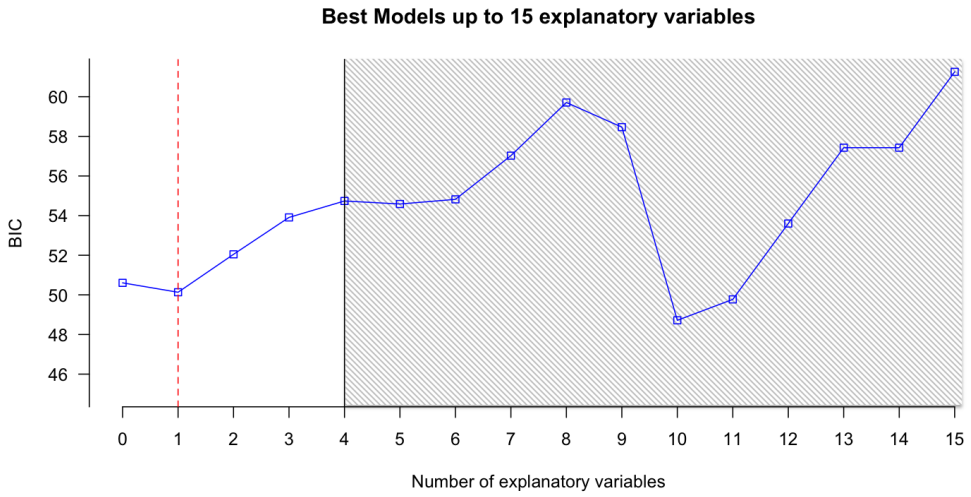
Tabela 3.14: Dez melhores modelos de mortalidade considerando toda a pesquisa exaustiva e respetiva significância de variáveis de acordo com o teste de Wald.

| Model order | Weight ($\times 10^{-3}$) | GA | MJD Inborn Delivery | MV | O2 | Surfactant | ETT Resuscitation | Maternal Age | Constant | BIC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | -5.65** | -0.82** | -2.78* | – – – | – – – | – – – | – – – | – – – | 27.47** | 115.47 |
| 4 | -5.59** | -0.74* | -2.88* | – – – | – – – | – – – | 1.12 | – – – | 24.80** | 116.27 |
| 4 | -5.68** | -0.88** | -2.74* | 16.91 | – – – | – – – | – – – | – – – | 12.29 | 116.59 |
| 4 | -5.33** | -0.92** | -2.97* | – – – | 2.83 | – – – | – – – | – – – | 27.34** | 116.61 |
| 4 | -5.54** | -0.86** | -2.65* | – – – | – – – | 1.96 | – – – | – – – | 26.41** | 116.72 |
| 5 | -5.242** | -0.87** | -3.22** | – – – | 2.67 | – – – | 1.17 | – – – | 25.44** | 117.12 |
| 5 | -5.20** | -0.98** | -2.85* | – – – | 2.98 | 2.13 | – – – | – – – | 26.50** | 117.38 |
| 5 | -5.356** | -1.01** | -2.94* | 17.09 | 2.96 | – – – | – – – | – – – | 12.24 | 117.40 |
| 3 | -7.53** | – – – | -2.89* | – – – | – – – | – – – | 1.32* | – – – | 7.48** | 117.42 |
| 6 | -5.46** | -1.03** | -3.73** | – – – | 3.27 | – – – | 1.53* | 0.10 | 26.27** | 117.79 |

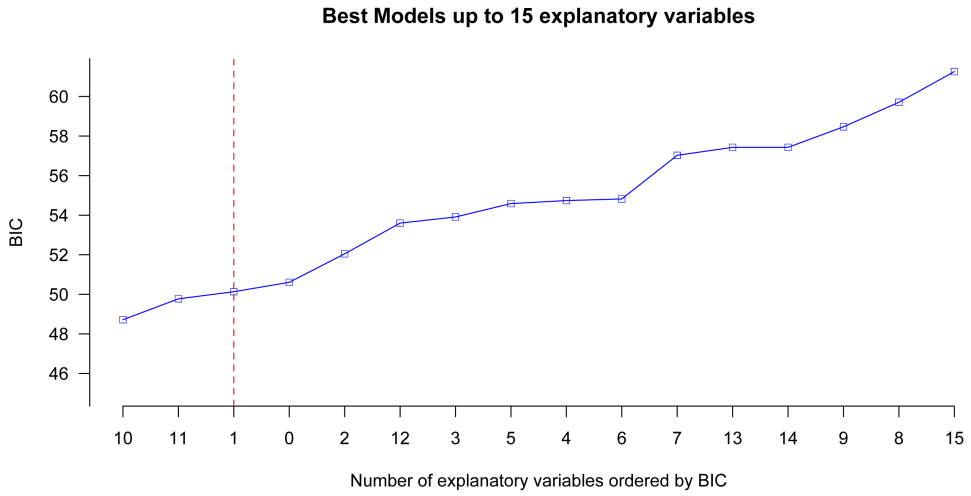
**p-value<0.01, *p-value<0.05;Reference class for MJD Inborn Delivery is other institutions and for binary variables yes/no is no

A construção de um modelo adequado capaz de descrever dois anos de morbilidade foi realizada seguindo todas as abordagens e etapas consideradas no caso da mortalidade.
Estamos cientes que neste caso, devido à reduzida dimensão da amostra, se torne mais árduo encontrar modelos credíveis para cada ordem $j$, $j \in \{0, ..., p\}$. De acordo com os cálculos efetuadas na subsecção 3.1.2, deverão ser adotados modelos até 1 variável, podendo no entanto esta escolha ser relaxada.

A figura 3.25 ilustra os resultados obtidos, considerando o valor de BIC para cada modelo.




(a)




(b)

Figura 3.25: Valor do critério BIC para cada ordem do modelo de morbilidade: (a) de acordo com a ordem do modelo (b) ordenado por ordem crescente de valor de BIC.

A partir do melhor modelo de ordem 4, inclusive, os modelos não conseguem convergir (Figura 3.25a), talvez devido ao elevado número de variáveis quando comparado com a dimensão das

observações. Uma ordenação crescente dos valores de BIC, permite-nos observar rapidamente a ordem dos melhores modelos (Figura 3.25b), concluindo-se que os dois melhores são os constituídos por 10 e 11 variáveis. Não obstante, a nossa procura cinge-se a modelos em torno de 1 variável e considerando no máximo a inclusão de 4 variáveis, pelo que a inspeção de ordens muito superiores não terão interesse.

Na tabela 3.16 (página 87) encontram-se detalhadas as variáveis presentes nos modelos em estudo (melhores modelos de ordem 1,2 e 3), assim como a significância das mesmas. Saliente-se que todas as variáveis apresentam coeficientes positivos, significando que os recém nascidos prematuros extremos que possuem essa característica têm maior probabilidade de possuírem anomalias futuras. Relativamente à significância das variáveis, os modelos de ordem 2 e 3 mostraram ser os melhores modelos dessas ordens, mesmo que nenhuma das variáveis seja significativa. Com estes resultados, não há muitas alternativas de escolha, sendo que o modelo que possui apenas uma variável significativa (Multifetal Gestation), será considerado para prever dois anos de morbilidade.
A equação a ele associada é a seguinte:

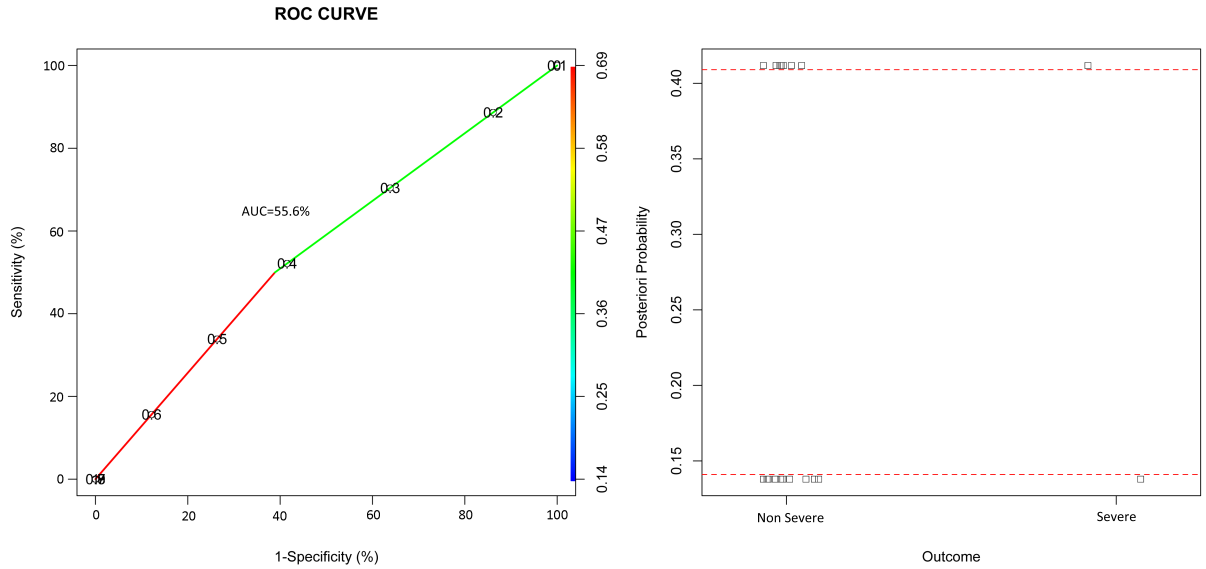
$$logit(p_1) = -1.83 + 1.48 \quad \text{Multifetal Gestation} \tag{3.36}$$

Face à reduzida escolha de modelos credíveis, foram também avaliados os dez melhores modelos até 3 variáveis decorrentes de toda a pesquisa exaustiva. Esta limitação no número de variáveis deve-se ao facto de não ser conseguida convergência para modelos de ordem superior a 3.
Na tabela 3.17 (página 87) encontram-se sintetizados os resultados obtidos.
Do panorama geral transparece que os dez melhores modelos de toda a pesquisa exaustiva possuem no máximo 2 variáveis. No entanto, a maioria destes não apresentam qualquer variável significativa. Esta situação só é invertida, apenas em dois modelos, nos quais a variável Mutlifetal Gestation é a única significativa. Um deles, corresponde ao melhor modelo que se obteve anteriormente (equação (3.36)) e o segundo corresponde ao décimo melhor modelo desta pesquisa exaustiva, constituído pelas variáveis Multifetal Gestation e BPD.

Perante esta situação, o modelo para prever dois anos de morbilidade será então o descrito pela equação (3.36). Não se conseguiu efetuar a avaliação de ajustamento aos dados deste modelo através do teste de $H\&L$, devido à limitação de este teste se basear na avaliação de decis de probabilidade. Neste caso, como o modelo apresenta unicamente uma variável dicotómica, apenas dois valores de probabilidade serão estimadas, não sendo portanto o teste aplicável (consultar subsecção 2.2.2). No que diz respeito à proporção de variância explicada, as estatísticas de Cox&Snell $R^2$ e Nagelkerke's $R^2$ apresentaram valores de 0.09% e 13.4% respetivamente.

A avaliação deste modelo, quanto ao carácter preditivo, mostra que não é capaz de fazer uma discriminação assertiva entre as classes Severe e Non Severe, AUC=55.6% (Figura 3.26a). Uma vez que este modelo só depende de uma variável, as probabilidades *a posteriori* estimadas apenas apresentam dois valores (0.1379310 e 0.4117647), consoante a Multifetal Gestation assumir o valor 0 ou o valor 1. Neste seguimento, não faz sentido a escolha de um ponto de corte ótimo, pois todo o ponto de corte que verifique a condição, $0.1379310 < cutoff < 0.4117647$, apresentará a mesma classificação. Além disto, as discretizações na curva ROC não são percetíveis, pois apenas traduzem a união de dois pontos de corte. Este fenómeno pode ser observado na figura 3.26b, onde os indivíduos da classe Non Severe se concentram em torno da probabilidade mais baixa e os

da classe Severe em probabilidades mais altas, embora havendo exceções em ambos os casos. As linhas a tracejado vermelho indicam os pontos de corte limite para uma classificação idêntica.




(a) (b)

Figura 3.26: Curva ROC para o estudo de um ponto de corte ótimo referente ao modelo de morbilidade.

O resultado de previsão deste modelo, encontra-se na tabela 3.15. A precisão foi de 60.0%, apresentando uma *Sensitivity* de 50.0% e *Specificity* de 61.1%. Apesar da percentagem de classificações corretas da classe Severe ser de 50% é necessário relembrar que a representatividade desta classe é muito baixa.

Tabela 3.15: Tabela de classificação referente a dois anos de morbilidade para um ponto de corte no intervalo ]0.1379310,0.4117647[, considerando o método da regressão logística.

| | | Observed | | Overall |
|---|---|---|---|---|
| | | **Severe** | **Non Severe** | Accuracy(%) |
| Predicted | **Severe** | 1 | 7 | |
| | **Non Severe** | 1 | 11 | |
| Accuracy(%) | | 50.0 | 61.1 | 60.0 |

Já para o ponto de corte usual, a classificação correta foi de 90%, uma vez que todos os bebés foram previstos como Non Severe. Consequentemente, atingiu-se os 0% de *Sensitivity* e o valor máximo (100%) de *Specificity*.

Tabela 3.16: Modelos de morbilidade e respetiva significância de variáveis de acordo com o teste de Wald.

| Model order | Multifetal Gestation | ETT Resuscitation | PVL | Surfactant | Constant |
|---|---|---|---|---|---|
| 1 | 1.48* | – – – | – – – | – – – | -1.83** |
| 2 | 1.24 | 1.16 | – – – | – – – | -2.51** |
| 3 | 1.05 | – – – | 36.08 | 17.01 | -18.57 |

**p-value<0.01, *p-value<0.05; Reference class for binary variables yes/no is no

Tabela 3.17: Dez melhores modelos de morbilidade até 3 variáveis considerando toda a pesquisa exaustiva e respetiva significância de variáveis de acordo com o teste de Wald.

| Model order | Multifetal Gestation | ETT Resuscitation | PVL | Surfactant | HMD | BPD | Constant | BIC |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.48* | – – – | – – – | – – – | – – – | – – – | -1.83** | 50.13 |
| 0 | – – – | – – – | – – – | – – – | – – – | – – – | -1.16** | 50.61 |
| 1 | – – – | 1.45 | – – – | – – – | – – – | – – – | -2.14** | 50.99 |
| 1 | – – – | – – – | 17.82 | – – – | – – – | – – – | -1.25** | 51.50 |
| 2 | 1.24 | 1.16 | – – – | – – – | – – – | – – – | -2.51 | 52.05 |
| 2 | – – – | – – – | 37.13 | 17.50 | – – – | – – – | -18.57 | 52.06 |
| 2 | 1.32 | – – – | 17.08 | – – – | – – – | – – – | -1.83** | 52.10 |
| 1 | – – – | – – – | – – – | – – – | 16.53 | – – – | -17.57 | 52.13 |
| 2 | – – – | 1.33 | 17.38 | – – – | – – – | – – – | -2.14** | 52.54 |
| 2 | 1.62* | – – – | – – – | – – – | – – – | 1.28 | -2.99* | 52.55 |

**p-value<0.01, *p-value<0.05; Reference class for binary variables yes/no is no

## 3.4 Discussão técnica dos resultados

Na presente secção pretendem-se comparar tecnicamente os resultados obtidos em ambos os métodos utilizados. Adicionalmente, procura-se também discutir algumas vantagens e desvantagens de cada um deles.

As medidas essenciais para a avaliação dos modelos obtidos para mortalidade e morbilidade encontram-se reportadas na tabela 3.18.

Tendo em conta o poder discriminante dos modelos, verifica-se em ambos que a regressão logística tem uma ligeira vantagem comparativamente com as árvores de classificação, dado que apresenta AUC superiores. Esta ligeira superioridade da regressão logística quando comparada com árvores de decisão é narrada em alguns artigos da literatura que reportam aplicações na área médica (Colombet et al., 2000; Austin et al., 2010). Todavia, outros autores afirmam que a diferença entre os dois métodos é muito ténue e, portanto, nenhum deles se supera relativamente ao outro (Kitsantas et al., 2006).

No nosso estudo, os resultados entre os métodos são bastante próximos tornando-se difícil escolher qual o mais adequado para classificar novos indivíduos.

No caso da mortalidade (Tabela 3.18a), os modelos obtidos por ambos os métodos são equivalentes, uma vez que identificaram as mesmas variáveis no modelo final: GA, Weight e MJD Inborn Delivery. No entanto, os modelos apresentam previsões diferentes dependentes do método utilizado e da maneira como este faz o ajuste. Neste contexto, verifica-se que a precisão global e as taxas de verdadeiros positivos (*Sensitivity*) e verdadeiros negativos (*Specificity*) conseguem ser superiores no método da regressão logística.

Algumas vantagens/desvantagens são apontados a ambos os métodos no que concerne à sua capacidade de aplicação e interpretação. Por exemplo, as árvores de classificação apresentam uma estrutura muito intuitiva permitindo reconhecer imediatamente a importância das variáveis nela contidas (Colombet et al., 2000; Camdeviren et al., 2007). Além do mais, as variáveis consideradas como fatores de risco são obtidas com mais detalhe, uma vez que é necessário determinar o valor a segmentar cada nó. No caso da mortalidade, verificou-se que um ponto de corte considerando 25 semanas de gestação, seria o limiar ideal para uma previsão correta. Assim, as árvores de classificação tornam-se vantajosas na identificação de grupos de indivíduos com características semelhantes, no entanto, ao contrário da regressão logística não permitem compreender qual a relação entre as variáveis explicativas e a variável preditora (Kitsantas et al., 2006).
A questão da identificação de subgrupos pode ser justificada nos modelos de mortalidade pelo facto de embora ambos os métodos tenham retornado as mesmas 3 variáveis como mais promissoras de mortalidade, as árvores de regressão obtiveram-nas sem que fosse necessário remover as variáveis mais promissoras de morbilidade, ou seja, identificaram grupos de indivíduos semelhantes de acordo com tais variáveis. Já a regressão logística necessitou da eliminação de algumas variáveis, tendo posteriormente chegado aos mesmos fatores de risco, através do método de pesquisa exaustiva. A regressão logística é considerada um método bastante poderoso quando combinado com métodos de seleção de variáveis (Austin et al., 2010).

Uma outra vantagem da regressão logística deve-se ao resultado que esta retorna. A possibilidade de serem estimadas as probabilidades *a posteriori* de cada indivíduo pertencer a uma classe, permite que seja estudada a relação do evento quando se tem em conta as diversas covariáveis (Worth and

Cronin, 2003; Kitsantas et al., 2006). O resultado das árvores cinge-se a que todos os indivíduos presentes num nó sejam classificados na mesma classe, pois o método atribui a classe de imediato.

Contudo, ao contrário da regressão logística, as árvores de classificação caracterizam-se como um método não paramétrico, onde não é assumida nenhuma suposição da distribuição de qualquer tipo de variáveis (covariáveis ou resposta), não sendo portanto necessário especificar a natureza dos dados (Worth and Cronin, 2003; Kitsantas et al., 2006).

Tabela 3.18: Desempenho dos modelos de mortalidade e morbilidade em ambos os métodos estudados.

(a)

| | Mortality | |
|---|---|---|
| | Classification Tree | Logistic Regression |
| Overall Accuracy | 69.4% | 70.0% |
| Sensitivity | 64.3% | 65.2% |
| Specificity | 76.2% | 76.5% |
| AUC | 70.2% | 78.6% |

(b)

| | Morbidity | |
|---|---|---|
| | Classification Tree | Logistic Regression |
| Overall Accuracy | 90.9% | 60.0% |
| Sensitivity | 0% | 50.0% |
| Specificity | 100% | 61.1% |
| AUC | 50.0% | 55.6% |

Os modelos de mortalidade parecem então ser muito semelhantes, quer nas medidas de avaliação, quer nas variáveis contidas nos modelos em ambos os métodos.
No caso do estudo de morbilidade, a regressão logística assume ter um poder discriminante superior ao das árvores de classificação (AUC=55.6% *versus* 50.0%), embora não permita mesmo assim discriminar completamente as duas classes (Table 3.18b). Já a precisão global assume maior valor no método das árvores de classificação (90.9%), no entanto, este modelo refere-se ao modelo nulo, dado que não foi conseguido determinar uma árvore de classificação para morbilidade e consequentemente, todos os bebés foram previstos na mesma classe. A regressão logística conseguiu identificar a variável Multifetal Gestation como possível preditora de morbilidade, sendo mesmo assim um modelo com capacidade preditiva bastante baixa. O facto de os modelos baseados em regressão logística de acordo com o tamanho da amostra deverem conter no máximo uma variável, poderá ter prejudicado. Acontece que o critério adotado para a seleção do melhor modelo, critério BIC, atribui uma penalização ao número de parâmetros. Sendo assim, variáveis categóricas com mais de dois níveis serão à partida excluídas. Por exemplo, a variável IVH possui três categorias e é mencionada na literatura como sendo das variáveis mais promissoras de morbilidade.

Nesta fase, é também importante relembrar que os modelos de morbilidade tiveram em conta amostras de treino e teste em que as representatividades das classes eram muito heterogéneas, podendo ser esta uma possível justificação para a baixa capacidade preditiva.

De acordo com os resultados obtidos e com uma análise sintetizada das vantagens e desvantagens destes métodos, entendemos que as árvores de classificação e a regressão logística poderão ser consideradas como complementares, partilhando da opinião semelhante de Kitsantas et al. (2006).

Em resumo, a regressão logística permite a obtenção de uma noção de risco e as árvores permitem hierarquizar variáveis.

A obtenção de um modelo preditivo aplicável na prática clínica não deve ser analisado apenas tecnicamente. As variáveis nele contidas e a respetiva relevância clínica que tais variáveis transmitem são consideradas fundamentais. Assim, procederemos à análise e comparação destes modelos do ponto de vista clínico no capítulo 4.

## 3.5 Conclusões

O método de regressão logística demonstrou, em ambas as previsões de mortalidade e morbilidade, uma ligeira vantagem relativamente à árvore de classificação.
Na previsão de um ano de mortalidade, a idade gestacional, o peso e o nascimento na MJD foram os preditores obtidos em ambos os métodos, sendo que os modelos mostraram uma precisão global de classificação correta de cerca de 70% e uma discriminação entre classe aceitável. Já no caso da morbilidade, não foi encontrada uma árvore de classificação significativa e o método de regressão logística destacou o fator de risco Multifetal Gestation. O desempenho deste último modelo foi reduzido mas no entanto mais favorável do que o desempenho da árvore de classificação, assumindo valores de 60% para a precisão global e 55.6% de AUC, revelando não fazer uma discriminação evidente entre classes.

O estudo de árvores de classificação baseadas em custos de má classificação permitiu compreender qual a gravidade de comunicação do perito de saúde no prognóstico de um recém nascido prematuro extremo aos respetivos pais.

# Capítulo 4

# Limite de viabilidade

Normalmente, a análise estatística complementa-se com a interpretação contextual dos resultados obtidos. Este capítulo encontra-se dividido em três secções. Na secção 4.1 pretende-se efetuar uma abordagem relativamente aos fatores de risco (variáveis) identificados pelos diversos métodos neste trabalho académico. Adicionalmente, far-se-á uma discussão detalhada da relevância clínica destes fatores, fazendo o contraponto com o que tem vindo a ser referenciado na literatura.
A secção 4.1 encontra-se, por sua vez, dividida em duas partes. A primeira incide nos fatores de risco associados com mortalidade e a segunda incide nos fatores de risco associados com morbilidade. Em cada uma das partes, serão discutidos os factores de risco precoces (Capítulo 2) e também os obtidos pelos modelos preditivos de regressão logística e árvores de classificação (Capítulo 3). Finalmente, os resultados obtidos serão comparados com os apresentados em Sá et al. (2012b,c), referentes a estudos paralelos desenvolvidos no decorrer desta tese em colaboração com a equipa médica da MJD. Decorrente desta colaboração, está em curso a elaboração de um artigo resumindo os resultados, que será brevemente submetido para publicação num jornal científico (Sá et al., 2012a).
Por fim, na secção 4.2 é apresentada uma discussão relativa ao desempenho dos modelos bem como do impacto que estes terão na prática clínica.

## 4.1 Discussão contextualizada dos fatores de risco

Nesta secção apresenta-se uma discussão contextualizada dos fatores de risco identificados sendo também apresentada uma comparação dos resultados obtidos neste trabalho científico com os obtidos nos trabalhos de Sá et al. (2012b,c,a). Estes últimos referem-se a estudos paralelos desenvolvidos no decorrer desta tese em colaboração com a equipa médica da MJD. O objetivo primordial incidiu na construção de modelos de mortalidade e morbilidade, respetivamente, baseados em regressão logística multivariada e contendo variáveis selecionadas de acordo com a experiência dos clínicos envolvidos.
É essencial frisar que o estudo de Sá et al. (2012b,c,a) e o desenvolvido no âmbito desta dissertação são bastante distintos. Por um lado, não se procedeu à divisão da amostra em conjunto de treino e de teste, tendo sido utilizada a amostra completa para construção e avaliação da capacidade preditiva do modelo. Por outro lado, as variáveis dos modelos foram indicadas por experiência clínica, não tendo havido qualquer tentativa de redução do número de variáveis do modelo ou inclusive a utilização de procedimentos do tipo stepwise para encontrar um modelo contendo apenas variáveis significativas.

### 4.1.1 Mortalidade

Numa primeira abordagem a este trabalho de investigação, procurámos identificar os fatores de risco mais precoces associados a um ano de mortalidade (Capítulo 2, subsecção 2.4.1). Para tal, foram selecionadas variáveis recolhidas, na sua maioria, no primeiro dia de vida. Estas variáveis foram agrupadas em subconjuntos de acordo com a sua "natureza", distinguindo fatores protocolares e fatores intrínsecos.
No que concerne aos **fatores protocolares**, uma análise baseada em regressão logística univariada assevera que o nascimento na MJD (Inborn delivery at MJD), o uso de corticoides (Antenatal Steroids), o parto por cesariana (Caesarean Delivery) e a não necessidade de intubação (ETT Resuscitation) diminuem significativamente o risco de mortalidade de um recém nascido prematuro extremo. Quando ajustados à idade gestacional (GA), apenas Inborn delivery at MJD permanece fator de risco significativo. Por fim, a regressão logística multivariada das intervenções protocolares demonstrou que Inborn delivery at MJD, não necessidade de ETT Resuscitation e o aumento da GA estão associadas a um baixo risco de mortalidade.
Relativamente aos **fatores intrínsecos**, através da regressão logística simples constatou-se que o género masculino (Male Gender), 5-min Apgar>3, o aumento da GA e o aumento do Peso (Weight) contribuem para a diminuição do risco de mortalidade. Depois de ajustados à GA, 5-min Apgar> 3 deixa de ser significativo e a rutura prolongada de membrana (MR>24 h) é um fator de risco significativo de baixa mortalidade quando comparada com MR<12 h. Finalmente, a análise conjunta demonstrou que apenas o aumento do Weight e da GA são fatores de risco significativos para a diminuição do risco de mortalidade.

Uma abordagem semelhante à descrita anteriormente encontra-se publicada no artigo The Express Group Members (2010). Os seus autores tiveram como objetivo analisar a associação entre as diversas intervenções clínicas e o risco de mortalidade no primeiro ano de vida em recém nascidos prematuros extremos nascidos na Suécia no período de 2004 a 2007. Note-se que, embora o estudo seja equivalente (com recurso à regressão logística simples, ajustada e multivariada), este incidiu exclusivamente em fatores considerados protocolares. Nesse estudo, o uso de corticoides (Antenatal Steroids), surfactante (Surfactant) e o nascimento num hospital classificado de alto nível (que corresponde à variável Inborn Delivery at MJD, para a realidade portuguesa) destacaram-se como estando associados a um baixo risco de mortalidade, quer na análise individual quer na ajustada

à GA. Já Caesarean Delivery apenas evidenciou estar associada a um aumento de sobrevivência na análise de regressão simples. No modelo multivariado, o uso de Antenal Steroids e Surfactant permanecem significativos e, em contraponto, o nascimento num hospital especializado deixa de ser significativo.

Os nossos resultados, a par do artigo anteriormente referenciado permitem que seja efetuada uma comparação entre a realidade Portuguesa e a realidade Sueca, constatando-se que existem fatores de risco protocolares que são sistematicamente identificados: corticoides (Antenatal Steroids), parto por cesariana (Caesarean delivery) e local de nascimento privilegiado (Inborn at MJD). Apesar de neste trabalho de investigação, o Surfactant ter sido incluído na análise, não foi identificado como fator de risco (subsecção 2.4.1). Esta questão poderá advir da forma como as variáveis são codificadas. Neste caso, a variável apenas nos indica se um recém nascido prematuro extremo usufruiu ou não de surfactante enquanto que no estudo sueco a variável em análise corresponde à utilização/não utilização de surfactante apenas nas primeiras duas horas após o parto. Esta diferença temporal na variável Surfactant poderá ser de facto decisiva, não sendo possível a comparação deste fator de risco entre os dois estudos.

Posteriormente, foram desenvolvidos os modelos preditivos baseados em regressão logística multivariada e árvores de classificação, sendo que ambos resultaram em modelos equivalentes, assumindo que as variáveis GA, Weight e MJD Inborn Delivery são as mais adequadas para prever mortalidade (Figura 4.1). Neste trabalho, o peso e a idade gestacional mantiveram-se simultaneamente significativos em todas as análises consideradas, corroborando o possível compromisso entre estas duas variáveis na definição do limite de viabilidade conforme sustentado por Seri and Evans (2008).

| Model order | | OR(95% CI) |
|---|---|---|
| Weight | $-5.65\times10^{-3}$** | 0.9944(0.9906,0.9981) |
| GA | -0.82** | 0.44(0.24,0.8) |
| MJD Inborn Delivey | -2.78* | 0.06(0.01,0.56) |
| Constant | 27.47** | − − − |

**p-value<0.01, p-value<0.05; Reference class for MJD Inborn Delivery is other institutions

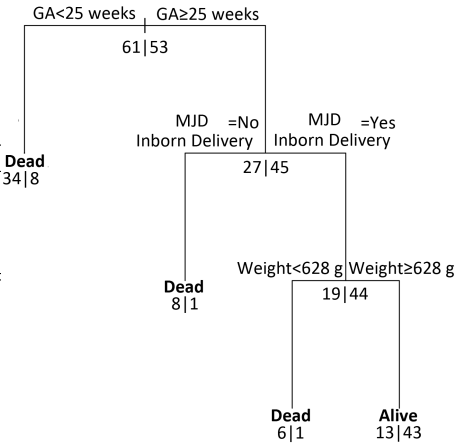



Figura 4.1: Modelo preditivo de mortalidade baseado em regressão logística (esquerda) e em árvores de classificação (direita).

A regressão logística permite uma análise dos coeficientes estimados do modelo. Neste caso, os coeficientes transmitem a relação entre o risco de mortalidade inerente a cada variável (quando as restantes são consideradas constantes) e são interpretados de acordo com o sinal que apresentam. Todas as variáveis mostram coeficientes com sinal negativo, indicando que tais variáveis diminuem o risco de mortalidade. Uma forma mais intuitiva para interpretação destes coeficientes, será obter os OR e os respetivos intervalos de confiança para cada variável, uma vez que estes permitem compreender a magnitude de cada variável preditora sobre o evento (mortalidade).

- GA: OR=0.44; este valor indica que por cada aumento de uma semana de gestação, o risco de um recém nascido prematuro morrer diminui 0.44 vezes.

- Weight: OR=0.9944; a interpretação é análoga à anterior.
- MJD Inborn Delivery: OR=0.06; conclui-se que o risco de mortalidade diminui em 0.06 se o recém nascido prematuro nascer na MJD, em comparação com os que são submetidos ao transporte após o nascimento.

As árvores de classificação têm como caraterística hierarquizar as variáveis e atribuir pontos de cisão fulcrais nas mesmas. Além disto, assumem regras de decisão que são imediatamente conclusivas, neste caso:

- Se GA<25 weeks → Dead
- Se GA≥ 25 weeks e MJD Inborn Delivery=No → Dead
- Se GA≥ 25 weeks e MJD Inborn Delivery=Yes e Weight< 628g → Dead
- Se GA≥ 25 weeks e MJD Inborn Delivery=Yes e Weight≥ 628g → Alive

Por exemplo, a variável GA e o ponto de corte de 25 semanas evidenciam ser das decisões mais importantes para descrever mortalidade.

**Aumento da Idade Gestacional diminui o risco de mortalidade**
A idade gestacional é sem dúvida uma das variáveis mais importantes para descrever mortalidade, sendo reportado na literatura que o risco de morte diminui à medida que a idade gestacional de um recém nascido prematuro extremo avança (Boussicault et al., 2012; The Express Group Members, 2010).
Inúmeros artigos na área médica baseiam-se em análises de regressão logística multivariada para a deteção de fatores associados a mortalidade. Nas investigação de Arad et al. (2008) e Boussicault et al. (2012), a variável GA é assumida como contínua, apresentando um OR de, respetivamente, 0.70 (0.53,0.93) e 0.55 (0.39,0.91), cujos valores são equivalentes aos encontrados neste estudo (Figura 4.1). Embora Arad et al. (2008) investigue recém nascidos de baixo Peso, a média das GA destes é inferior a 26 semanas, podendo portanto ser um trabalho comparável.
Noutros estudos, a variável GA é categorizada, como é o caso de Bacak et al. (2005) e Boussicault et al. (2012), onde o risco de mortalidade é significativamente mais elevado para idades gestacionais inferiores a 24 semanas, quando comparadas com idades gestacionais mais elevadas (<27 semanas), sendo que os OR são tipicamente superiores a 2.

Curiosamente, no modelo preditivo de mortalidade baseado em árvores de classificação, o ponto de cisão obtido para a GA foi 25 weeks (Figura 4.1), estando em sintonia com a literatura médica. Adicionalmente, o estudo de Ambalavanan et al. (2006) baseado em árvores de classificação para prever conjuntamente mortalidade e morbilidade de recém nascidos de baixo peso, corrobora com os nossos resultados ao referir um maior risco de mortalidade para idades gestacionais abaixo das 25 semanas de gestação.

Estudos realizados nos Estados Unidos por Kaempf et al. (2009); Lantos and Meadow (2009) sustentam que os recém nascidos com idades gestacionais entre as 23 e as 24 semanas (GA<25 weeks) têm poucas probabilidades de sobreviver, pelo que se recomenda que estes prematuros não sejam assistidos na unidade de cuidados intensivos neonatais. Kaempf et al. (2009) justificam esta questão, alegando que a principal causa desta recomendação passa por não ser compensatório tal assistência, na medida em que o desfecho desta situação resultante em óbito do recém nascido prematuro ou mesmo lesões neurológicas significativas atinge um valor superior a 50%.

Outra justificação adicional decorre da experiência de vários clínicos que defendem não ser digna a reanimação de crianças que possam vir a ter problemas severos a longo prazo, ou cujos clínicos não consigam ajudar os pais a resolver tais complicações. Já para recém nascidos prematuros extremos com idades gestacionais iguais ou superiores a 25 semanas (GA$\geq$25 weeks), recomenda-se que seja disponível todo o investimento tanto na sala de parto como na unidade de cuidados intensivos, em concordância com outro estudo americano reportado por Seri and Evans (2008)

Em Portugal, o panorama é bastante semelhante. Torna-se óbvio que quando estamos perante situações extremas e indiciadoras de mau prognóstico, as decisões a tomar podem não ser muito difíceis. No entanto, o que se torna verdadeiramente árduo é definir quais serão as situações limite (Silva and Carvalho, 2008).
A Secção de Neonatologia da Sociedade Portuguesa de Pediatria (Peixoto et al., 2004) referiu nos Consensos Nacionais em Neonatologia diversas recomendações a ter em conta quando se trata de recém nascidos prematuros extremos, sendo que toda a atuação clínica depende da idade gestacional que o recém nascido apresenta. A existência deste tipo de protocolos é considerada imprescindível na orientação das tomadas de decisão, segundo a discussão de diversas opiniões clínicas apresentadas por Silva (2008).
Peixoto et al. (2004) referem que na realidade Portuguesa, os recém nascidos prematuros extremos podem ser divididos em apenas três grupos:

- Os que a maioria dos clínicos concordam que não devem ser assistidos: GA$<$ 24 weeks
- Os que devem ser tratados, segundo grande parte das opiniões médicas: GA$\geq$ 25 weeks
- Os que são suscetíveis de dúvidas e algumas divergências sobre o tratamento: 24$\leq$GA$<$25 weeks

Neste contexto, o critério das 24 semanas traduz o limite em que é sabido que se torna dispensável efetuar qualquer tratamento ao recém nascido prematuro. Porém, existe uma incerteza sobre recém nascidos que nascem entre as 24 e as 25 semanas de gestação. Este dúvida está de acordo com os artigos mencionados, referindo que as 25 semanas são de facto um limiar influente na determinação de mortalidade.
Note-se que a resolução da variável GA é reportada em unidades de semanas. No entanto, seria útil adotar-se uma resolução em dias, pois poderá fazer toda a diferença nas tomadas de decisão.

**Aumento do Peso diminui o risco de mortalidade**
O peso de um recém nascido prematuro é também um fator importante para descrever mortalidade de um recém nascido prematuro extremo. É intuitivo afirmar que quanto maior for o peso de um recém nascido prematuro extremo, maior é a probabilidade de este sobreviver. Esta associação foi encontrada nos modelos preditivos de mortalidade baseados em regressão logística multivariada e em árvores de decisão.

A regressão logística multivariada indicou um OR=0.9944 (0.9906,0.9981), com o peso codificado em unidades de grama (Figura 4.1). Alguma bibliografia refere que o peso está associado significativamente com a mortalidade embora, apenas em análises univariadas, como é o caso de Arad et al. (2008). Por outro lado, a investigação de Boussicault et al. (2012) indica numa análise univariada, que o peso não altera a mortalidade (OR=0.99 (0.99,1.00)), considerando também a resolução de 1 grama. A amostra considerada em Boussicault et al. (2012) comporta 108 observações, sendo portanto uma dimensão próxima à da amostra utilizada nesta tese. O facto de terem considerado o peso como não influente na mortalidade, poderá ser consequência de arredondamento, pois o

intervalo de confiança é muito próximo de 1. Caso procedêssemos a um arredondamento nos extremos do intervalo de confiança do nosso resultado, a conclusão seria a mesma de Boussicault et al. (2012).

Na árvore de classificação, é possível verificar um ponto de cisão de 680 g para a variável Weight, quando a idade gestacional é superior a 25 semanas (Figura 4.1). Seri and Evans (2008) retratam estudos acerca do limite de viabilidade em função do peso, onde referem que recém nascidos prematuros com Weight $<$ 500 g são muito imaturos e a probabilidade de sobrevivência é muito diminuta. Contrariamente, recém nascidos prematuros com Weight$\geq$ 600 g têm uma probabilidade de sobrevida bastante maior. Esta não complementaridade de pesos ($<$500 g e $\geq$600 g) apresenta uma resolução de 100 gramas, correspondendo aos casos que são considerados "limite de viabilidade"e, onde os procedimentos a adotar não se encontram estabelecidos. Também de acordo com o trabalho de Seri and Evans (2008), estes aclaram que com base na informação disponível na literatura e em estudos americanos recentes, as crianças com GA$\geq$25 weeks e Weight$\geq$600 g têm uma probabilidade de sobrevivência de aproximadamente, 60% a 70%. Este intervalo percentual coincide com os resultados deste trabalho, dado que 60% dos recém nascidos prematuros extremos, nas condições anteriormente referidas, sobrevivem ao primeiro ano de vida.

**Nascer num hospital especializado diminui o risco de mortalidade**
Por fim, mas não menos importante, o local de nascimento torna-se um fator essencial. Nascer num hospital considerado especializado e de grande qualidade é uma mais valia para qualquer recém nascido, sobretudo para os recém nascidos prematuros extremos, uma vez que podem usufruir rapidamente de todos os recursos disponíveis e necessários para um tratamento adequado (The Express Group Members, 2010).
Além do mais, o tempo de intervenção é um fator precioso e bastante decisivo no desfecho destes recém nascidos, uma vez que os bebés que necessitam de transporte entre hospitais, poderão ser afetados negativamente (Arad et al., 2008). Para efetuar este tipo de transporte é importante que os recém nascidos sejam acompanhados por especialistas e também por equipamentos imprescindíveis, de forma a que sejam garantidos os serviços mínimos essenciais (Araújo et al., 2011). Contudo, dada a fragilidade destes bebés, estes serviços prestados nos transportes poderão ser insuficientes, sendo preferível o transporte *in-utero*.

Nesta investigação, todos os recém nascidos prematuros extremos foram seguidos na MJD, alguns nasceram na MJD e outros chegaram à MJD transferidos de outros hospitais (corresponde à variável MJD Inborn Delivery).

Estudos realizados nos Estados Unidos, em França e no Canadá indicam que o transporte agrava o prognóstico de morte de um recém nascido prematuro, apresentando em análises multivariadas de regressão logística OR de 2.2(1.8,2.6), de 3.32(1.02,10.9) e de 1.5(1.1,2.0), respetivamente (Boussicault et al., 2012; Bacak et al., 2005; Sankaran et al., 2002).

Também nesta pesquisa, o nascimento na MJD (e a não necessidade de transporte) foi identificado como um fator associado a um menor risco de mortalidade em todas as análises realizadas, estando portanto em concordância com estudos internacionais.

**Estudo paralelo em colaboração com a equipa médica da MJD**
No trabalho de Sá et al. (2012b), foi criado um modelo de mortalidade baseado em regressão

logística multivariada, contendo variáveis selecionadas[1], de acordo com a experiência dos clínicos envolvidos. Nesta abordagem destacaram-se como variáveis significativas associadas a um decréscimo de mortalidade, o aumento da GA (OR=0.353(0.208,0.599)), o aumento do Weight (OR=0.996(0.993, 0.999)) e o uso de Antenatal Steroids (OR=0.150(0.004,0.510)). Já a MR<12h quando comparada com MR>24h mostrou aumentar o risco de morte de um recém nascido prematuro extremo (OR=3.876(1.406,10.680)).

Neste sentido, fatores de risco como GA, Weight, Antenatal Steroids e MR são novamente obtidos, em concordância com os resultados desta tese. Para além disto, artigos recentes reportam que além da idade gestacional e do peso, existem também outros fatores importantes para prever mortalidade, como por exemplo, os corticoides (Tyson et al., 2008; Medlock et al., 2011) e a rutura de membrana (Blumenfeld et al., 2010). Estes artigos apresentam na maioria um tamanho da amostra muito superior ao deste trabalho paralelo (183 observações): Tyson et al. (2008) apresentam resultados com base numa amostra de 3000 indivíduos, enquanto Medlock et al. (2011) consideram amostras entre 59 e cerca de 12000 observações, uma vez que se trata de um artigo onde se comparam diversos estudos. Também, Blumenfeld et al. (2010) apresentam resultados sobre uma amostra de cerca de 17000 observações. No entanto, as conclusões dos estudos são idênticas às obtidas nesta tese e às obtidas no trabalho paralelo, indicando evidência estatística para uma amostra de muito menor dimensão.

### 4.1.2 Morbilidade

Se para o estudo de mortalidade a dimensão da amostra era reduzida ($n$=163), este problema agrava-se para o estudo da morbilidade ($n$=74), uma vez que as observações no estudo de morbilidade são necessariamente as observações no estudo de mortalidade que sobrevivem. O tamanho da amostra muito reduzido poderá traduzir-se numa limitação do estudo, em particular na circunstância de não se encontrar evidências estatísticas e modelos significativos com uma amostra de dimensão tão pequena.

De facto, não foram identificados fatores de risco de morbilidade precoces, quer intrínsecos quer protocolares.
Em geral, os intervalos de confiança associados aos OR estimados por regressão logística simples e ajustada apresentam grande amplitude, à exceção do Weight (OR=1.00(0.996,1.006); OR=0.999(0.991,1.005)) e da Maternal Age (OR=1.02(0.92,1.14); OR=1.01(0.9,1.13) ), sugerindo que um aumento da dimensão da amostra poderia conduzir à identificação destes fatores de risco. Uma circunstância semelhante é também o fator de risco Multifetal Gestation que apresenta OR na análise individual e ajustada à GA de 3.48(0.99,12.23) e 2.94(0.76,11.26), respetivamente. Embora os intervalos de confiança sejam amplos, estes estão muito próximos de deixar de conter o valor 1. Todavia, o modelo multivariado para descrever dois anos de morbilidade é constituído pelo fator de risco Multifetal gestation, mesmo sem que este seja significativo. Como referido na subsecção 2.4.2, este resultado poderá ser consequência de uma limitação do algoritmo do procedimento stepwise, em que se preferiu ter um modelo multivariado com uma variável com $p - value = 0.052$ no teste de Wald ao modelo nulo.

A procura exaustiva de um modelo de regressão logística para previsão de morbilidade tornou-se complexa devido ao reduzido número de observações. De acordo com o critério de seleção utilizado (subsecção 3.3.1) o melhor modelo para prever morbilidade aos dois anos de idade

[1]Foram consideradas as variáveis GA, Weight, Gender, PS, Multifetal Gestation, Antenatal Steroids, MR, Caesarean Delivery e Iatrogenic Delivery.

consta apenas da variável Multifetal Gestation (Figura 4.2). O OR e o respetivo intervalo de confiança associado a esta variável é de 4.37 (1.05,18.29), estimando que o risco de desenvolvimento neurológico com sequelas é 4.37 vezes superior se a gestação for múltipla, em comparação com uma gestação simples. Este modelo é equivalente ao modelo multivariado obtido por stepwise para os fatores intrínsecos, mas conduziram a conclusões diferentes considerando uma significância de 5%: num caso Multifetal Gestation é fator de risco significativo e no outro caso não. Este facto é consequência não só do tamanho da amostra, mas também da grande variabilidade inerente aos dados, que conduz a conclusões diferentes dependendo da amostra que se considere para estimação dos coeficientes. Estes resultados apontam para o facto de não ser possível encontrar neste estudo factores significativos associados a morbilidade.

| Model order | | OR(95% CI) |
|---|---|---|
| Multifetal Gestation | 1.48* | 4.37(1.05,18.29) |
| Constant | -1.83** | – – – |

**p-value<0.01, *p-value<0.05; Reference class for binary variables yes/no is no

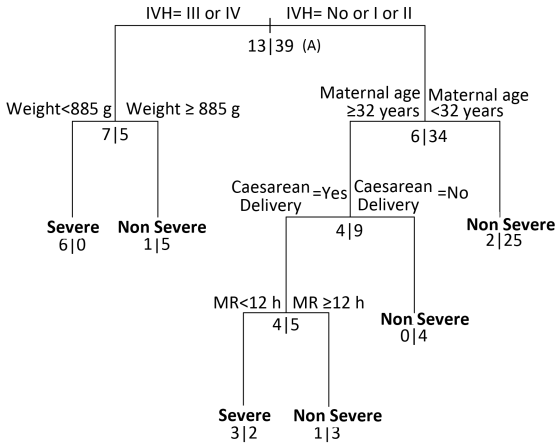



Figura 4.2: Modelo preditivo de morbilidade baseado em regressão logística (esquerda) e em árvores de classificação (direita).

As árvores de classificação também não permitiram obter modelos significativos. Foi apenas identificada a árvore máxima, uma vez que o processo de poda indicou que os fatores de risco incluídos não revelavam importância em termos de previsão. A não existência de uma árvore podada é mais uma evidência de que o tamanho da amostra poderá ser reduzido e a variabilidade dos dados poderá ser grande. A árvore máxima apresenta como fatores de risco IVH, Weight, Maternal age, Caesarean Delivery e MR (Figura 4.2), parecendo estar em concordância com a experiência dos clínicos da MJD. A exceção são as variáveis Maternal age e Caeserean Delivery, não havendo também suporte científico na literatura a indicar estas duas variáveis como fatores de risco de morbilidade. Serão necessários, portanto outros estudos para comprovar estes fatores como determinantes de morbilidade.
A árvore máxima indica as regras de decisão em cada nó. Em particular, as regras apresentadas de seguida permitiram obter nós puros, isto é, não demonstraram dúvidas na determinação do desfecho de morbilidade:

- Se IVH=III or IV e Weight<885 g → Severe
- Se IVH=No or I or II e Maternal age ≥32 years e Caeserean Delivery=No → Non Severe

De seguida discute-se a relevância clínica dos factores de risco encontrados, embora não significativos.

**Gestação Múltipla aumenta o risco de morbilidade**
No que refere à gestação múltipla, é consensual que esta agrava o prognóstico do desenvolvimento

neurológico de um recém nascido prematuro extremo, tendo esta situação sido reportada recentemente num artigo de revisão suíço (Berger et al., 2011). Esta desvantagem é também encontrada num livro americano sobre Obstetrícia e Ginecologia escrito por Beckaman et al. (2010), onde os autores referem que gravidezes múltiplas estão associadas a um aumento do risco de morbilidade, de 3 a 4 vezes, quando comparadas com gestações simples. A investigação de Wang et al. (2011) baseada numa análise de regressão logística univariada, revelou que a gestação múltipla foi dos únicos fatores associados a um aumento do risco de morbilidade (OR=5.2(1.6,17.0)), para além de GA<26 weeks e de baixo peso (Weight<800g), para um $p\text{-}value<0.05$.
Estes resultados estão de acordo com o descrito pelo modelo de regressão logística deste trabalho, que determina um OR um pouco superior a 4 (Figura 4.2), à semelhança dos indicados anteriormente.

Existem vários fatores que contribuem para o aumento da morbilidade nas gravidezes gemelares, tais como o aumento da incidência de parto pré-termo, rutura prematura de membranas, síndrome de transfusão feto-fetal, anomalias congénitas, alterações do crescimento, compressão e procedência do cordão umbilical, descolamento prematuro da placenta e complicações decorrentes do próprio trabalho de parto (da Graça, 2010). Estes fatores contribuem para o aumento da incidência de intercorrências no período neonatal (síndrome de dificuldade respiratória, hemorragias intraventriculares, sépsis, etc...) que podem vir a originar défices neurológicos como cegueira, surdez ou paralisia cerebral.

**Hemorragia intraventricular grave aumenta o risco de morbilidade**
McCrea and Ment (2008) frisam que apesar de vários estudos reportarem que o desfecho cognitivo poderá estar associado diretamente à idade gestacional da criança (GA), dados recentes indicam que existem preditores mais importantes para o *outcome* neurológico, enfatizando a hemorragia intraventricular. De facto, estudos recentes consideram IVH como um fator de risco de elevada importância para o desenvolvimento neurológico adverso (Boussicault et al., 2012; Peralta-Carcelen et al., 2009; Zeitlin et al., 2008).

A IVH é geralmente reportada de acordo com a sua gravidade na progressão do desenvolvimento clínico. Assim, consideram-se hemorragias intraventriculares graves as IVH de grau III e IV, sendo que as principais complicações que dela advêm são a paralisia cerebral e problemas mentais (Peralta-Carcelen et al., 2009).

É claro que o agravamento do desenvolvimento neurológico de um recém nascido prematuro depende muito da sua maturidade. É neste sentido que McCrea and Ment (2008) indicam que a gravidade da hemorragia intraventricular assume-se inversamente proporcional à idade gestacional e ao peso. Neste contexto, recém nascidos prematuros extremos com idades gestacionais pequenas e/ou baixo peso à nascença são mais suscetíveis a hemorragias graves e, portanto, maior risco de morbilidade.
No entanto, neste trabalho não foi encontrada associação significativa entre IVH e GA e entre IVH e Weight (teste Kruskal-Wallis, $p - value = 0.4542$ e $p - value = 0.2364$ respetivamente), para corroborar o trabalho de McCrea and Ment (2008).

**Aumento do Peso diminui o risco de morbilidade**
O peso é considerado uma variável de extrema relevância para o desenvolvimento de recém nascidos prematuros. Quanto mais pesado for um recém nascido, mais favorável poderá ser o seu prognóstico, na medida em que este é mais maduro e mais tolerante a intervenções médicas.

O estudo de Tyson et al. (2008) reporta que um modelo multivariado contendo o peso, a idade gestacional e mais alguns fatores prevê com maior precisão morbilidade do que um modelo univariado considerando apenas a GA, indicando que o peso é um fator de risco importante de morbilidade mas não individualmente.

Na árvore de classificação deste trabalho é também notório o compromisso entre Weight e IVH (Figura 4.2), condizente com o que foi descrito no fator de risco anterior. Caso um recém nascido tenha uma hemorragia intraventricular considerada grave, o peso torna-se um fator determinante na decisão do outcome. Também o estudo de Ambalavanan et al. (2006) referente à construção de uma árvore de classificação que descreve simultâneamente mortalidade e morbilidade, encontrou a variável peso como promissora do evento.

**Rutura prolongada de Membranas diminui o risco de morbilidade**
A rutura de membrana é considerada uma causa de complicações da prematuridade extrema a longo prazo (Waters and Mercer, 2009), nomeadamente problemas sequentes a infeções por parte da mãe ou do recém nascido prematuro.

De forma a combater tais complicações de desenvolvimento neurológico futuro, tem-se optado por administrar corticoides enquanto as membranas permanecem rotas e até à altura do parto. Assim sendo, a rutura prolongada de membranas permite uma maior administração de corticoides, contribuindo para que o recém nascido prematuro adquira uma maior maturidade pulmonar derivada da administração de corticoides por mais tempo, o que poderá evitar problemas futuros.

A árvore de classificação obtida (Figura 4.2) indica que a duração da rutura de membrana é decisiva para o desfecho da morbilidade, em particular MR<12h conduz a diagnóstico de morbilidade severa. Este resultado está de acordo com o trabalho de Waters and Mercer (2009), que indica que a administração de corticoides no caso de ruturas de membrana sensivelmente próximas da ocorrência do parto poderá não ser suficiente para assegurar a viabilidade fetal.

Estudos recentes indicam mesmo que em recém nascidos prematuros extremos que sofreram de ruturas de membranas e tiveram administração de tal fármaco, o risco de desenvolvimento neurológico adverso é bastante inferior, em comparação com os que não usufruíram de corticoides (Carlo et al., 2011).

**O Parto por cesariana como indicador de uma diminuição de morbilidade permanece em estudo**
A via de parto mais indicada para um recém nascido prematuro extremo é alvo de alguma controvérsia na literatura médica. No caso da mortalidade, em algumas situações existe consenso de que o parto por cesariana nas idades gestacionais <27 semanas pode ser indicador de uma diminuição a mortalidade (subsecção 2.4.1). No caso da morbilidade, esta questão ainda não está esclarecida.

Um artigo de revisão recentemente publicado por Berger et al. (2011) indica que, apesar do aumento do número de cesarianas em prematuros extremos, não se verificou uma redução de morbilidade, pelo que não há evidências a sustentar que esta via de parto deva ser realizada de forma rotineira baseada apenas na idade gestacional. Também Wang et al. (2011) referem diversas opiniões publicadas na literatura médica acerca do tipo de parto: alguns autores indicam benefícios da cesariana relativamente à diminuição da taxa de morbilidade, enquanto outros expõem opiniões contrárias. Não obstante, na investigação destes autores, o parto por cesariana não

mostrou benefícios de morbilidade. Neste contexto, o tema da melhor via de parto permanece em aberto, uma vez que estas recomendações se baseiam apenas em estudos retrospetivos sem evidência estatística e em opiniões de peritos.

Neste trabalho, a árvore de classificação indica que o parto vaginal (Caeserean Delivery=No) se associa a recém nascidos prematuros extremos sem sequelas futuras, indicando estar concordante com o descrito anteriormente.

**Estudo paralelo em colaboração com a equipa médica da MJD**
No contexto da morbilidade, foram criados dois modelos. No primeiro pretendeu-se prever morbilidade de acordo com as variáveis escolhidas[2] (Sá et al., 2012b). Numa segunda abordagem, foram acrescentadas duas novas variáveis ao modelo de morbilidade anterior: Weight e GA (Sá et al., 2012c). Em ambos os casos a única variável significativa foi a IVH. O primeiro modelo sugeriu que Non IVH e IVH=I ou II estão associadas a uma diminuição do risco de sequelas quando comparadas com IVH de graus III ou IV. No segundo modelo, apenas se verificou que Non IVH está associada a um menor risco de morbilidade quando comparada com IVH grave (graus III ou IV). No entanto, a hemorragia de nível intermédio (graus I ou II) não mostrou ser significativa quando se adicionou o peso e a idade gestacional, indicando que nos casos de gravidade intermédia de IVH, o peso e a idade gestacional poderão ser fatores relevantes no desfecho de morbilidade.

## 4.2 Desempenho dos modelos preditivos e impacto na prática clínica

Após a construção de um modelo preditivo é imprescindível fazer-se uma avaliação quanto ao desempenho dos mesmos. Recentemente, Medlock et al. (2011) fizeram uma revisão sistemática de diversos artigos presentes na literatura que reportam a modelos preditivos de mortalidade em recém nascidos prematuros. Esta revisão relata as principais diretrizes encontradas nos estudos, como os tipos de variáveis utilizadas e os métodos mais frequentes. Além disto, estes autores destacam a importância de serem utilizadas medidas de avaliação que possam ser comparáveis entre estudos e que sejam consistentes com os objetivos a atingir, referindo como as mais utilizadas: o teste de H&L (para o caso da regressão logística, subsecção 2.2.2), a AUC (subsecção 3.1.4) e a precisão global de classificações corretas (subsecção 3.1.4, equação 3.6).

As árvores de classificação e os modelos de regressão logística, para mortalidade e morbilidade, obtidos neste trabalho de investigação foram comparados anteriormente na secção 3.4. Resumindo, a regressão logística mostra ter uma ligeira vantagem, relativamente às árvores de classificação, nomeadamente:

- **Mortalidade**: o modelo de regressão logística apresenta uma AUC de 78.6% e precisão de 70.0%; a árvore de classificação assume uma AUC de 70.2% e 69.4% de precisão global.
- **Morbilidade**: o modelo de regressão logística tem uma AUC de 55.6% e uma precisão de 60.0%; a árvore de classificação apresenta uma AUC de 50.0% e precisão de 90.9%.

Para o caso da **mortalidade**, os modelos parecem fazer uma boa discriminação das classes a prever (AUC$\geq$70%) assim como uma classificação correta global consideravelmente elevada. Estes resultados são concordantes com a literatura, em que modelos preditivos de mortalidade baseados em regressão logística multivariada apresentam uma AUC tipicamente entre 74% e 80%(Tyson

[2]Integram este estudo as variáveis BDP, IVH, PVL, HMD, PDA, Infection, Antenatal Steroids e Vaginal Delivery.

et al., 2008; Medlock et al., 2011).

Já no que diz respeito à **morbilidade**, o poder discriminante dos modelos é muito pobre (AUC≈ 50%). Também na literatura, modelos de morbilidade baseados em regressão logística indicam uma AUC de cerca de 74%, quando considerados modelos multivariados que englobem gestação múltipla, peso, idade gestacional, entre outros fatores (Tyson et al., 2008), o que não foi observado nesta tese. No entanto, caso se avaliassem apenas as precisões, concluir-se-ia que os modelos deste trabalho de investigação seriam modelos muito adequados. Por esta razão, é de extrema importância a utilização de mais do que uma medida de avaliação de modelos.

Neste trabalho académico encontraram-se valores de 69% e 90% de precisão para as árvores de mortalidade e morbilidade, respetivamente. Estando cientes da limitação da precisão global da árvore de morbilidade, os resultados obtidos para mortalidade mostram ser semelhantes aos do artigo de Ambalavanan et al. (2006) que reporta valores de precisão entre os 61% e os 62%, para uma árvore de mortalidade e morbilidade conjunta.

Os modelos preditivos baseados em regressão logística desenvolvidos neste estudo serão agora comparados com os apresentados no trabalho conjunto de Sá et al. (2012b,c). A tabela 4.1 apresenta medidas de ajuste referentes a cada modelo em estudo, como a significância do teste de H&L, medidas de variância explicada (Cox&Snell $R^2$ e Nagelkerk's $R^2$) e capacidade preditiva avaliada através da precisão global de classificações corretas (Overall Accuracy) para um ponto de corte de 0.5, para que os modelos sejam comparáveis entre si (Fawcett, 2006). Os modelos obtidos nos diversos trabalhos são bastante distintos nas variáveis (consultar notas de rodapé).

Tabela 4.1: Comparação dos modelos de regressão logística de mortalidade e morbilidade obtidos nesta investigação com os obtidos no estudo paralelo de Sá et al. (2012b) e Sá et al. (2012c).

| | Mortality | | Morbidity | | |
|---|---|---|---|---|---|
| | This work[1] | Sá et al. (2012b) [2] | This work [3] | Sá et al. (2012b) [4] | Sá et al. (2012c) [5] |
| p-value of HL test | 0.1697 | 0.194 | — | 0.289 | 0.763 |
| Cox& Snell $R^2$ | 33.4% | 39.1% | 0.09% | 27.0% | 26.8% |
| Nagelkerk's $R^2$ | 44.6% | 52.6% | 13.4% | 42.4% | 41.9% |
| Overall Accuracy* | 67.5% | 78.1% | 90% | 87.8% | 85.7% |

* The cutoff value is 0.5 for all models, including those developed in this thesis

Pela análise da tabela constata-se que todos os modelos são significativos e ajustam-se adequa-

[1]Modelo de mortalidade obtido por pesquisa exaustiva (Figura 4.1, esquerda). O modelo contém as variáveis GA, Weight e MJD Inborn Delivery.

[2]Modelo de mortalidade que contém variáveis selecionadas por experiência clínica: GA, Weight, Gender, PS, Multifetal Gestation, Antenatal Steroids, MR, Caesarean Delivery e Iatrogenic Delivery, onde apenas GA, Weight, Antenatal Steroids e MR são significativas.

[3]Modelo de morbilidade obtido por pesquisa exaustiva (Figura 4.2, esquerda). O modelo contém apenas a variável Multifetal Gestation.

[4]Modelo de morbilidade que contém variáveis selecionadas por experiência clínica: BDP, IVH, PVL, HMD, PDA, Infection, Antenatal Steroids e Vaginal Delivery, onde apenas IVH é significativa.

[5]Modelo de morbilidade que adiciona duas novas variáveis ao modelo anterior, Weight e GA. Apenas IVH demonstrou ser significativa.

damente aos dados (H&L $p - value > 0.05$ para todos os casos). A exceção é o modelo de morbilidade desenvolvido nesta tese, o qual não foi possível ser avaliado pelo teste de H&L. Esta limitação advém do fato de que este teste é baseado na comparação de decis da probabilidade estimada e, como o modelo obtido é constituído por uma só variável dicotómica, só apresenta dois valores de probabilidade possíveis (Subsecção 3.3.2 , equação 3.36).

No que diz respeito à percentagem de variância explicada, os modelos de mortalidade explicam um pouco mais em relação aos de morbilidade, até 53% em comparação com cerca de 40%. Estes valores indicam que os modelos conseguem explicar alguma variância de mortalidade/morbilidade.

Quanto à capacidade preditiva, o modelo de mortalidade em Sá et al. (2012b) traduz uma precisão global de 78.1%, enquanto que o obtido neste trabalho apresenta 67.5%. Os modelos diferem apenas em cerca de 10% na precisão global, o que poderá sugerir que os modelos são equivalentes, tendo em conta as possíveis limitações inerentes a esta comparação. Eventualmente, será preferível ter um modelo com menos variáveis, como é o do obtido nesta investigação, em relação a um modelo com 9 variáveis das quais apenas 4 se mostraram associadas significativamente com a mortalidade. Os modelos identificam duas variáveis significativas em comum, Weight e GA, o que corrobora a importância clínica destas variáveis (Peixoto et al., 2004).
No caso da morbilidade, a capacidade preditiva do modelo obtido neste trabalho é de 90%. Esta percentagem poderia traduzir grande precisão do modelo, no entanto, é consequência do facto das probabilidades estimadas serem todas inferiores a 0.5, isto é, todos os indivíduos são previstos na mesma classe. Os outros dois modelos (Sá et al., 2012b,c) apresentam uma capacidade preditiva equivalente (87.8% e 85.7%) de onde se conclui que a inclusão do peso e da idade gestacional não acrescentou benefícios à precisão do modelo.

É difícil declarar qual a percentagem de precisão considerada razoável para um modelo de previsão. Segundo Ambalavanan et al. (2006), um grau de predição de cerca de 60.0% não é suficiente para se tomarem decisões clínicas. Neste trabalho, quer os modelos de árvores de classificação, quer os modelos de regressão logística apresentam uma classificação correta global geralmente superior a 60.0%. Assim, os modelos de previsão resultantes deste trabalho, para além de explicarem adequadamente mortalidade ao um ano e o desenvolvimento neurológico a longo prazo, estão também corroborantes com a prática clínica, valendo a pena serem utilizados. Não obstante, o uso destes deverá ser limitado e não abusivo, uma vez que há necessidade de se procurarem constantemente modelos melhores que consigam traduzir AUC, variância explicada e capacidades preditivas mais elevadas.

Estes modelos preditivos são um contributo indispensável para a sobrevivência de qualidade de um recém nascido prematuro extremo, identificando atempadamente eventuais medidas a longo prazo e mudanças na prática clínica. Além disto, devido às incertezas que advêm de idades gestacionais muito baixas, o envolvimento dos pais nos processos de decisão é fundamental, pelo que estes modelos permitem, ainda melhor informação, traduzindo-se num contributo efetivo para o aconselhamento parental.

# Capítulo 5

# Conclusões

Este trabalho teve como objetivo a identificação de fatores de risco precoces e tardios associados a mortalidade (ao 1º ano de vida) e a morbilidade (avaliada ao 2º ano de vida) de recém nascidos extremamente prematuros portugueses (<27 semanas de gestação). Adicionalmente, foram construídos modelos preditivos, baseados em regressão logística e em árvores de decisão com o intuito de prever o desfecho de mortalidade e morbilidade.

Os resultados deste trabalho para a realidade portuguesa corroboram a elevada importância do fator de risco idade gestacional (do inglês gestational age, GA) para a definição do *limite de viabilidade* (Peixoto et al., 2004). Esta variável tem ainda a vantagem de ser recolhida facilmente e de forma não invasiva, não causando transtorno nem à criança nem à mãe. No entanto, ainda que significativa, a GA apresenta uma associação moderada com o desfecho, pelo que um prognóstico clínico baseado unicamente nesta variável poderá ser muito limitado. Esta restrição motiva a procura de outros fatores de risco complementares à GA (Tyson et al., 2008) e a construção de modelos preditivos mais informativos.

No caso da mortalidade, a GA, o peso e o nascimento em hospital menos diferenciado com necessidade de transferência para tratamento, foram identificados como fatores de risco significativos, quer com regressão logística quer com árvores de classificação. O desempenho dos dois modelos foi estimado com base nas mesmas amostras de treino e teste, indicando um desempenho ligeiramente superior para o modelo de regressão logística (overall accuracy 70.0% *versus* 69.4% e AUC de 78.6% *versus* 70.2%). Já para o caso da morbilidade, não foram identificados fatores de risco precoces. O modelo de regressão logística identificou o fator de risco gestação múltipla e não foi encontrada uma árvore de classificação significativa. A regressão logística teve um desempenho estimado moderado (overall accuracy 60.0% e AUC de 55.6%), ligeiramente mais expressivo do que um modelo nulo.

Neste estudo, houve várias variáveis precoces que não foram identificadas como fatores de risco significativos de mortalidade e/ou morbilidade, por falta de evidência estatística. Nestes casos, o teste estatístico não foi conclusivo, isto é, não rejeitou a hipótese nula dos coeficientes associados serem nulos, o que poderá ter sido consequência de uma limitação do tamanho da amostra. Um eventual aumento ao tamanho da amostra (supondo o desvio padrão inalterável) traduzir-se-ia em intervalos de confiança de menor amplitude e na possível identificação de mais fatores de risco significativos. As conclusões do estudo dos modelos preditivos foram também condicionadas pela dimensão da amostra, essencialmente por dois motivos. Por um lado, o tamanho máximo dos modelos foi ajustado ao tamanho da amostra para garantir uma estimação adequada dos seus

parâmetros, o que condicionou o número máximo de fatores de risco identificados como relevantes. Por outro lado, o tamanho da amostra condicionou a evidência estatística de ser fator de risco. No caso da regressão logística, o impacto incidiu na maior amplitude dos intervalos de confiança sobre os coeficientes e, no caso das árvores de classificação, incidiu na menor expressividade da diferença do número de observações em cada classe num nó terminal e consequente poda desse nó. Um exemplo da limitação do tamanho da amostra foi o caso da variável gestação múltipla nas abordagens por regressão logística para o estudo de morbilidade. Por um lado, não houve evidência estatística a 5% para considerar a gestação múltipla como fator de risco intrínseco (p-value=0.052, teste de Wald) e, por outro, o modelo preditivo com uma única variável identificou a gestação múltipla de todas as variáveis da base de dados (p-value=0.043, teste de Wald). Houve uma alteração nos resultados obtidos, possivelmente devido ao facto de se terem utilizado amostras diferentes para a estimação dos coeficientes. Foi realizada uma experiência por *bootstrap* para averiguar a questão da significância desta variável (Capítulo 2). Finalmente, o aumento do tamanho da amostra também poderia beneficiar o estudo, permitindo inclusive uma maior separação das classes sim/não na mortalidade, por exemplo, possibilitando uma análise comparativa distinguindo as caraterísticas daqueles que não morrem, dos que morrem mais cedo e dos que morrem mais tarde (classes Early Neonatal Death, Late Neonatal Death e Postneonatal Death da figura 1.1).

Houve também limitação ao nível da codificação das variáveis utilizadas neste estudo. Dando um exemplo concreto, a variável idade gestacional deveria ser recolhida em unidades de dia em vez de em unidades de semana. Esta maior resolução na escala de medida poderia permitir a definição do *limite de viabilidade* com maior precisão, respondendo melhor à questão do que fazer com crianças entre as 24 e as 25 semanas de gestação. Rennie (1996) refere que no *limite de viabilidade*, a hipótese de sobreviver aumenta cerca de 2% por dia. Assim sendo, com a resolução da variável GA em unidades de semanas, ao aumento de uma semana de GA corresponderá um aumento de 14% de hipóteses de sobrevivência, fazendo do intervalo de uma semana demasiado longo e justificando, assim, a necessidade de maior resolução na avaliação de GA. Geralmente, a idade gestacional é estimada com base na história menstrual corroborada por uma ecografia precoce (ultrasons), apresentando erros de medição que advêm da capacidade limitada do aparelho ecocardiográfico. Tyson et al. (2008) reportam que estimativas de idades gestacionais baseadas nestes exames de ultrasons podem conter um erro variando entre 4 a 14 dias, dependendo das semanas de gestação em que for realizada a ecografia. Assim sendo, há também uma precisão muito baixa na estimação da idade gestacional por este método, o que justifica inteiramente a necessidade de serem desenvolvidos métodos mais precisos na estimação da GA, uma vez que quão melhor for a estimativa de GA, melhor se poderá decidir as etapas ulteriores (Peixoto et al., 2004).
Idênticas melhorias na codificação de variáveis poderiam também ser sugeridas para as variáveis corticoides e infeção. Tipicamente além do sim/não, reportar também o número de intervenção, dose aplicada, número de dias, etc. Naturalmente que o acréscimo de informação na codificação destas variáveis terá de ser acompanhado pelo aumento do tamanho da amostra.

O facto de este trabalho ter incidido numa amostra referente a recém nascidos prematuros extremos seguidos exclusivamente na MJD, pode levar a que a extrapolação das conclusões inerentes a este estudo para a população portuguesa possa ser limitada. Os resultados obtidos talvez reflitam apenas uma realidade à escala regional, sendo necessário efetuar estudos a nível nacional para confirmar estas conclusões na população portuguesa.

Apesar dos resultados obtidos neste trabalho, há ainda um enorme esforço de investigação a prosseguir. Atualmente, os recursos disponíveis na MJD já permitem que estes recém nascidos prematuros sejam monitorizados continuamente através de parâmetros fisiológicos recolhidos em

tempo real como, por exemplo, a frequência cardíaca e parâmetros respiratórios. Na prática clínica, esta informação é utilizada para caraterizar o estado corrente do prematuro mas poderia também servir para prever a evolução e o desfecho desses recém nascidos. De facto, esta é a tendência atual de trabalhos científicos muito recentes e apresentados na $34^{th}$ *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*(2012), uma das conferências internacionais mais conceituadas na área da engenharia biomédica. Mikhno and Ennett (2012) e Precup et al. (2012) indicam que o grau de acoplagem entre sinais cardiovasculares e sinais respiratórios pode indicar o momento favorável para desintubar um recém nascido prematuro, minimizando a duração da ventilação mecânica e garantindo a não necessidade de reintubação, uma vez que este processo aumenta o risco de mortalidade. Índices correntemente desenvolvidos para estimar a sensibilidade do baroreflexo arterial também poderão ser utilizados para esse prognóstico (Gouveia, 2009). Zwanenburg et al. (2012) sugerem que a análise do eletroencefalograma é útil na avaliação do estado funcional e em eventuais lesões no cérebro de recém nascidos prematuros. Orlandi et al. (2012) propõem um sistema automático capaz de gravar o choro e os movimentos de um recém nascido prematuro, e estimar as frequências fundamentais do trato vocal que permitam detetar precocemente o distúrbio de autismo.
Todos estes trabalhos têm o objectivo de aumentar a informação disponível do estado dos recém nascidos extremamente prematuros com base na informação que já é monitorizada no dia-a-dia da unidade hospitalar, por forma a sugerir um prognóstico mais preciso sem alterar a rotina clínica.

Os modelos construídos neste trabalho, apresentaram-se significativos e capazes de prever adequadamente o desfecho de mortalidade/morbilidade. Todavia, é indispensável continuar a procurar outros modelos que possam desvendar de forma ainda mais explícita e concreta tais desfechos. Estes modelos são considerados uma ferramenta preciosa na previsão de mortalidade/morbilidade em recém nascidos prematuros extremos, muitos deles no *limite de viabilidade*, tornando-se úteis na orientação de tomadas de decisão dos profissionais de saúde, em tratamentos antecipados e no aconselhamento parental. Além disto, são uma mais valia também na prevenção de prognósticos, e, o facto de terem sido validados quanto à sua capacidade preditiva permitem que possam ser utilizados no futuro, contribuindo para uma diminuição das taxas de mortalidade e prevenção de desenvolvimento neurológico severo a longo prazo.

Neste contexto, é imperativo continuar a recolher informação de recém nascidos prematuros extremos, quer a nível local (MJD) quer nacional, para que estudos nesta área estejam em constante atualização, visando proporcionar um futuro mais otimista a estas crianças e aos pais.

# Capítulo 6

# Contributos científicos

Parte do trabalho de investigação descrito nesta tese resultou na apresentação do trabalho em quatro conferências mais vocacionadas para a área estatística.

1. **Januário, A.**, Gouveia, S., da Costa, J. P., Sá, M. I., Almeida, A., Carvalho, C., Saraiva, J. P., Fonte, M., and Soares, P. (2012). Risk factors associated with one-year mortality of extremely premature newborns in Portugal.
   In *Book of Abstracts of 5th Meeting of Young Researchers of University of Porto (IJUP'12),* page 118. Universidade do Porto. (Porto, fevereiro 2012)

2. **Januário, A.**, Gouveia, S., da Costa, J. P., Sá, M. I., Almeida, A., Carvalho, C., Saraiva, J. P., Fonte, M., and Soares, P. (2012). Logistic Regression in the study of one-year mortality in extremely premature newborns: impact of protocol and intrinsic risk factors.
   In *Book of Abstracts of XIX Jornadas de Classificação e Análise de Dados (JOCLAD2012),* pages 81-84. Associação Portuguesa de Classificação e Análise de Dados (CLAD)/ Instituto Politécnico de Tomar. (Tomar, março 2012)

3. **Januário, A.**, Gouveia, S., da Costa, J. P., Sá, M. I., Almeida, A., Carvalho, C., Saraiva, J. P., Fonte, M., and Soares, P. (2012). Prediction of one year mortality in extremely premature newborns using classification trees.
   In *Book of Abstracts of* $4^{as}$ *Jornadas de Iniciação à Investigação Clínica,* pages 125-126. Departamento de Ensino, Formação e Investigação do Centro Hospitalar do Porto (DEFI-CHP). (Porto, junho 2012)

4. **Januário, A.**, Gouveia, S., da Costa, J. P., Sá, M. I., Almeida, A., Carvalho, C., Saraiva, J. P., Fonte, M., and Soares, P. (2012). Prediction methods for mortality and morbidity prognostic of portuguese extremely premature newborns.
   In *Book of Abstracts of XX Congresso da Sociedade Portuguesa de Estatística*, pages 269-273. Sociedade Portuguesa de Estatística (SPE). (Porto, setembro 2012)

Adicionalmente, decorrente do estudo paralelo com a equipa médica da MJD, houve também apresentação de trabalho em duas conferências na área médica.

**5.** Sá, M. I., Fonte, M., Saraiva, J., Carvalho, C., Soares, P., **Januário, A.**, Gouveia, S., and Almeida, A. (2012). Management and outcome of extremely premature infants.
In *Book of Abstracts of XLI Jornadas Nacionais de Neonatologia da Sociedade Portuguesa de Pediatria,* pages 46-47. Secção de Neonatologia da Sociedade Portuguesa de Pediatria. (Braga, maio 2012)

**6.** Sá, M. I., Fonte, M., Saraiva, J., Carvalho, C., Soares, P., **Januário, A.**, Gouveia, S., and Almeida, A. (2012). Management and outcome of extremely premature infants.
In *Book of Abstracts of 17th World Congress on Controversies in Obstetrics, Gynecology & Infertility (COGI)*, page xx. Federação das Sociedades Portuguesas de Obstetrícia e Ginecologia (ESPOG).*Resumo submetido aceite.* (Lisboa, novembro 2012)

Estão ainda em preparação dois artigos para submissão a revistas científicas, um de índole estatística (**7**) e outro de índole médica (**8**):

**7. Januário, A**., Gouveia, S., da Costa, J. P., Sá, M. I., Almeida, A., Carvalho, C., Saraiva, J. P., Fonte, M., and Soares, P. (2012). Prediction methods for mortality and morbidity prognostic of portuguese extremely premature newborns. *to be submitted.*

**8.** Sá, M. I., Fonte, M., Carvalho, C., Soares, P., Almeida, A., **Januário, A.**, Gouveia, S., and Saraiva, J. (2012a). Premature infants under 27 weeks gestacional age: predicting outcome and improving parental counseling. *to be submitted.*

Nas páginas seguintes são apresentados os resumos submetidos a conferências/congressos. Adicionalmente, são também anexados posters referentes às comunicações em formato de painel.

# Risk factors associated with one-year mortality of extremely premature newborns in Portugal

**A. Januário [1,2], S. Gouveia[2], J. Pinto da Costa[1,2], M. I. Sá[3], A. Almeida[3], C. Carvalho[3], J. P. Saraiva[3], M. Fonte[3], P. Soares[3]**

[1] Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Portugal
[2] Gabinete de Estatística, Modelação e Aplicações Computacionais, CMUP, Portugal
[3] Unidade Maternidade de Júlio Dinis – Centro Hospitalar do Porto (MJD), Portugal

The decreasing mortality rate of premature newborns results from recent improvements in perinatal obstetric and neonatal care. This study aims to identify the risk factors associated with one-year mortality in Portuguese extremely premature newborns (<27 weeks of gestational age, when average is 40). The data was collected from 205 cases followed up at MJD from 2000 to 2009. **In this preliminary study**, the mortality odds ratio was estimated for each protocol related factor from binary logistic regression (Table 1, [1]). Simple regression **(a)** indicated that lower mortality is associated with use of Antenatal Steroids and non-Vaginal Delivery, in accordance with [2]. In Portugal, non-Intubation and Birth at MJD also diminished the risk of death. Because Gestational Age (GA) determines the protocol procedures, the OR was adjusted for GA **(b)** and, of all factors, only Birth at MJD was related with a lower mortality risk. Birth at MJD continues to be significant in the multivariate analysis **(c)**, besides non-Intubation and increasing GA with OR 0.37(0.25,0.55). This result indicates that even when adjusted for the use of protocol interventions and GA, birth at MJD is still associated with a reduced risk, implying that the survival advantage of birth at MJD was not associated with the studied factors and remains to be explained. Updated information on premature Portuguese newborns is expected from the results of this study, with inherent repercursion in clinical practice.

**Table 1:** Mortality odds ratio (OR) and 95% Confidence Intervals for each protocol related factor, where OR<1 indicates decrease risk of mortality.

| | **All n=167*** | **Dead n=87** | **(a) Crude** | **(b) Adjusted** | **(c) Multivariate** |
|---|---|---|---|---|---|
| Pregnancy Surveillance | 155 | 79 | **0.52**(0.15,1.8) | **1.12**(0.27,4.6) | ------- |
| Antenatal Steroids | 142 | 68 | **0.29**(0.11,0.77) | **0.42**(0.14,1.24) | ------- |
| Vaginal Delivery | 71 | 45 | **2.23**(1.19,4.17) | **1.15**(0.55,2.39) | ------- |
| Iatrogenic Delivery | 20 | 10 | **0.91**(0.36,2.31) | **1.13**(0.4,3.16) | ------- |
| Intubation | 119 | 68 | **2.04**(1.03,4.03) | **2.04**(0.96,4.33) | **2.27**(1.04,4.99) |
| Surfactant | 150 | 81 | **2.15**(0.76,6.12) | **2.34**(0.75,7.28) | ------- |
| Birth at MJD | 146 | 70 | **0.22**(0.07,0.68) | **0.23**(0.07,0.77) | **0.21**(0.06,0.71) |
| Epoch (2007-2009) | 36 | 21 | **1.22**(0.53,2.82) | **1.17**(0.45,3) | |
| (2003-2006) | 73 | 35 | **0.8**(0.4,1.6) | **0.85**(0.4,1.81) | ------- |
| (2000-2002) | 58 | 31 | **1**[Reference] | **1**[Reference] | |

Binary Logistic Regression: (a) simple; (b) adjusted for GA; (c) multivariate with stepwise procedure including GA; *=205 cases excluding stillbirths, delivery room death and missing values.

References:

[1] Hosmer, D. and Lemeshow, S. (2000), *Applied Logistic Regression*, John Wiley and Sons.

[2] The Express Group Members (2009), *One-Year Survival of Extremely Preterm Infants After Active Perinatal Care In Sweden,* JAMA, 301(21), 2225-2233.

XIX Jornadas de Classificação e Análise de Dados
Instituto Politécnico de Tomar, 28 – 31.Março.2012

# Logistic Regression in the study of one-year mortality in Portuguese extremely premature newborns: impact of protocol and intrinsic risk factors

**A. Januário[1], S. Gouveia[2], J. Pinto da Costa[3], M. I. Sá[4], A. Almeida[5], C. Carvalho[6], J. P. Saraiva[7], M. Fonte[8], P. Soares[9]**

*[1]Departamento de Matemática, Faculdade de Ciências da Universidade do Porto (FCUP); Gabinete de Estatística, Modelação e Aplicações Computacionais - Centro de Matemática da Universidade do Porto (GEMAC-CMUP), anafjanuario@hotmail.com;*
*[2]GEMAC-CMUP, sonia.gouveia@fc.up.pt;*
*[3]FCUP e CMUP, jpcosta@fc.up.pt;*
*[4]Serviço de Ginecologia e Obstetrícia, Unidade Maternidade Júlio Dinis - Centro Hospitalar do Porto (MJD-CHP), misabelrcsa@gmail.com;*
*[5]Serviço de Neonatologia, MJD-CHP, maria.alexandra.almeida@gmail.com;*
*[6]Serviço de Neonatologia, MJD-CHP, carmencarvalho01@gmail.com;*
*[7] Serviço de Ginecologia e Obstetrícia, MJD-CHP, saraivajp@hotmail.com;*
*[8]Serviço de Neonatologia, MJD-CHP, miguelfonte@gmail.com;*
*[9] Serviço de Neonatologia, MJD-CHP, mpccsoares@hotmail.com;*

## Summary

In recent years, the survival rate of extremely premature infants has increased, as a result of improved perinatal medical care (Bhaumik *et al.* (2004); The Express Group Members (2009)). These newborns are very small, and their organs are less developed than those born later and therefore, intrinsic factors can also determine their mortality. In this study, logistic regression is used to identify protocol related and intrinsic risk factors associated with one-year mortality of extremely premature newborns in Portugal.

**Keywords** Binary Logistic Regression, extremely premature newborns, mortality, neonatology.

## 1. Methods

Data were collected from 205 anonymous newborns followed up at MJD-CHP between 2000 and 2009, within a joint collaboration between Neonatal Intensive Care Unit and Obstetrics-Gynecology Departments at MJD-CHP, and approval by the Ethics Committee of CHP. The cases included extremely premature newborns with gestational age lower than 27 weeks (when most pregnancies last around 40 weeks). The dataset comprises several variables including maternal medical and previous obstetric history, data on pregnancy and delivery, infant condition at birth, selected neonatal procedures and one-year infant mortality (yes/no).

In this study, risk factors associated with one-year mortality were identified from binary

logistic regression (Hosmer *et al.,*2000). This analysis refers to the generalized linear model $y_i = g(\mu_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_q x_{qi}$ where $i = 1, \ldots, n$ is the case number, $q$ represents the number of risk factors and $g(\mu_i)$ is the link function (Hosmer *et al.,*2000). The variable $y_i$ can either be 0 or 1, where $y_i = 1$ indicates that $i^{th}$ newborn has died before one-year of life. The regression coefficients $\beta_j, j = 1, \ldots, q$, are estimated by conditional maximum likelihood and, from those, the odds ratio (OR) are obtained for each factor. For dichotomous factors (yes/no), the OR is the ratio of the odds of an event occurring in the "yes" group with respect to the "no" group (reference group). An OR=1 indicates that mortality is equally likely to occur in both groups, whereas an OR<1 indicates that mortality is more likely to occur in the reference group. For continuous variables, the OR is interpreted with respect to a unit increase in the variable, as there is no reference class.

The analysis was carried out in three steps. First, the ORs were estimated independently for each factor by means of simple logistic regression. Second, ORs were adjusted for gestational age (GA) in order to discard possible associations between the studied factors and clinical protocol or medical judgment, as it is known that both are determined by GA. Finally, the joint effect of several risk factors was studied by multivariate logistic regression. Goodness of fit for the multivariate models was assessed through the Hosmer-Lemeshow (HL) test, indicating a poor fit if the significance value is less than 0.05. Model quality was evaluated from the pseudo-R2 statistics Cox&Snell and Nagelkerke's R2 values (Hosmer *et al.,*2000), which are approximations of the proportion of variance of mortality explained by the model.

## 2. Results and Discussion

The risk factors were grouped in two subsets - *protocol* and *intrinsic* - according to their origin. On one hand, protocol related factors included e.g. pregnancy surveillance (Table 1), and were considered to evaluate the importance of clinical perinatal interventions. On the other hand, intrinsic factors such as gender and weight were also considered (Table 2).

With respect to protocol factors (**Table 1a**), simple logistic regression indicated that the use of antenatal steroids and caesarean delivery are associated with lower mortality, in accordance with (The Express Group Members, 2009). In the Portuguese population, unneeded endotracheal/ETT resuscitation, as judged by medical opinion, and inborn delivery at MJD also decreased the risk of death (with 95% confidence). After adjusting for GA only inborn delivery at MJD remained statistically significant (**Table 1b**). Finally, multivariate analysis of protocol interventions (**Table 1c**) indicate that unneeded ETT resuscitation, inborn delivery at MJD, and increasing GA are risk factors decreasing significantly mortality risk. This result is according to expected as ETT resuscitation is a medical intervention frequently used on newborns with most critical condition and, therefore, with few survival chances. The place of birth was also a

significant factor indicating that MJD inborn newborns may benefit from earlier standard of care despite the others that needed to be transferred from other institutions.

**Table 1:** Mortality OR for protocol related factors where OR<1 indicates decrease risk of mortality. The OR 95% Confidence Intervals are reported in parenthesis.

| | **All n=167*** | **Dead n=87** | **(a) Crude** | **(b) Adjusted** | **(c) Multivariate** |
|---|---|---|---|---|---|
| Pregnancy Surveillance | 155 | 79 | **0.52**(0.15,1.80) | **1.12**(0.27,4.60) | ------- |
| Antenatal Steroids | 142 | 68 | **0.29**(0.11,0.77) | **0.42**(0.14,1.24) | ------- |
| Caesarean Delivery | 96 | 42 | **0.45**(0.24,0.84) | **0.87**(0.42,1.81) | ------- |
| Iatrogenic Delivery | 20 | 10 | **0.91**(0.36,2.31) | **1.13**(0.40,3.16) | ------- |
| ETT Resuscitation | 119 | 68 | **2.04**(1.03,4.03) | **2.04**(0.96,4.33) | **2.27**(1.04,4.99) |
| Surfactant | 150 | 81 | **2.15**(0.76,6.12) | **2.34**(0.75,7.28) | ------- |
| Inborn Delivery at MJD | 146 | 70 | **0.22**(0.07,0.68) | **0.23**(0.07,0.77) | **0.21**(0.06,0.71) |
| Epoch (2007-2009) | 36 | 21 | **1.22**(0.53,2.82) | **1.17**(0.45,3.00) | |
| (2003-2006) | 73 | 35 | **0.80**(0.40,1.60) | **0.85**(0.40,1.81) | ------- |
| (2000-2002) | 58 | 31 | **1**[Reference] | **1**[Reference] | |
| GA (Weeks) | --- | --- | ------- | ------- | **0.37**(0.25,0.55) |

Binary Logistic Regression: (a) simple; (b) adjusted for GA; (c) multivariate with stepwise procedure including GA; *=205 cases excluding stillbirths, delivery room death and missing values.

A similar analysis was carried out for intrinsic variables. Simple logistic regression indicated that mortality risk decreases for male gender, 5-min Apgar>3, increasing weight and increasing GA (**Table 2a**). When adjusting for GA, 5-min Apgar≤3 was no longer significant while prolonged rupture of membranes (MR>24 hours before delivery) was revealed as a risk factor of lower mortality when compared to MR<12 hours (**Table 2b**), in accordance with (Blumenfeld *et al*., 2010). Finally, multivariate analysis pointed out weight and GA as the significant intrinsic risk factors (**Table 2c**). In this study, male gender was a protective risk factor when evaluated by simple regression, against literature reports (Bhaumik *et al.*, 2004). However, we found that male newborns have a significant higher median weight at birth than females (p-value 0.001, unilateral Mann-Whitney U test) and there are no gender differences between median GA (p-value 0.53, bilateral Mann-Whitney U test). This result is in accordance with the fact that, for the same GA, male newborns are heavier than females (Bhaumik *et al.*, 2004) and therefore, also justifies why gender was no longer a significant risk factor identified by multivariate analysis.

Both multivariate models (Table 1c and 2c) were found to describe adequately the data, as assessed by HL statistics (p-value 0.52 and 0.89, respectively). The value of Cox&Snell statistics were 22.9% (protocol) and 24.1% (intrinsic). The Nagelkerke's R2 is an adaptation of the Cox&Snell statistics with the advantage of ranging between 0 and 1. In this study, the corresponding values were 30.5% (protocol) and 32.1% (intrinsic), indicating that the major part of the mortality variance is yet to be explained. Therefore, these multivariate models (based on variables observed very early in time, almost after birth) are expected to provide a limited

impact for the purpose of predicting one-year mortality in extremely premature newborns.

**Table 2:** Mortality OR for intrinsic factors where OR<1 indicates decrease risk of mortality. The OR 95% Confidence Intervals are reported in parenthesis.

| | All n=153* | Dead n=81 | (a) Crude | (b) Adjusted | (c) Multivariate |
|---|---|---|---|---|---|
| Small for GA | 17 | 10 | **1.31**(0.47,3.64) | **2.26**(0.73,6.95) | ------- |
| Male gender | 89 | 40 | **0.46**(0.24,0.89) | **0.43**(0.20,0.89) | ------- |
| Primipara | 75 | 44 | **1.57**(0.83,2.98) | **1.37**(0.67,2.81) | ------- |
| Multifetal Gestation | 57 | 31 | **1.10**(0.57,2.12) | **1.28**(0.61,2.68) | ------- |
| 5-min Apgar ≤ 3 | 12 | 10 | **4.93**(1.04,23.31) | **3.66**(0.68,19.61) | ------- |
| MR ≥ 24h | 51 | 22 | **0.53**(0.27,1.07) | **0.42**(0.19,0.93) | |
| 12-24h | 10 | 5 | **0.70**(0.19,2.6) | **0.70**(0.17,2.89) | ------- |
| < 12h | 92 | 54 | **1**[Reference] | **1**[Reference] | |
| Weight (g) | ------- | ------- | **0.993**(0.991,0.996) | **0.996**(0.993,0.999) | **0.996**(0.993,0.999) |
| Maternal Age (Years) | ------- | ------- | **0.98**(0.93,1.04) | **1.01**(0.95,1.08) | ------- |
| GA (Weeks) | ------- | ------- | **0.33**(0.22,0.50) | ------- | **0.44** (0.28,0.70) |

Binary Logistic Regression: (a) simple; (b) adjusted for GA; (c) multivariate with stepwise procedure including GA; *=205 cases excluding stillbirths, delivery room death and missing values.

## 3. Conclusions

Several risk factors were significantly associated with one-year mortality of extremely premature newborns. Multivariate analysis considering protocol factors indicated that lower mortality was associated with unneeded ETT resuscitation, inborn delivery at MJD and increasing gestational age, while multivariate analysis based on intrinsic factors indicated increasing weight and increasing gestational age protective factors of the newborns against mortality. Considerable repercursion and impact in clinical practice is expected from the results of this study, as well as updated information about premature Portuguese newborns.

**Acknowledgement:** This work was partially supported by GEMAC-CMUP and CMUP, financed by FCT Portugal through POCI2010/POCTI/POSI programmes, with national and CSF funds. We also thank to Abbot Laboratórios for the financial support.

## References:

BHAUMIK, U. et al (2004) *Narrowing of Sex Differences in Infant Mortality in Massachusetts. J Perinatol*, 24, 94-99.

BLUMENFELD, Y.J. et al (2010) *The effect of preterm premature rupture of membranes on neonatal mortality rates*, *Obstet Gynecol*, 116(6), 1381-1386.

HOSMER, D.W., LEMESHOW, S. (2000) *Applied Logistic Regression*, Second Edition, John Wiley & Sons, Inc.

THE EXPRESS GROUP MEMBERS (2009) *One-Year Survival of Extremely Preterm Infants After Active Perinatal Care in Sweden*. *JAMA*, 301(21), 2225-2233.

Ref [3]. Resumo da apresentação em poster em 4as Jornadas de Iniciação à Investigação Clínica

**Prediction of One Year mortality in extremely premature newborns using classification trees**

Ana Januário
*Departamento de Matemática, Faculdade de Ciências da Universidade do Porto (FCUP) e Gabinete de Estatística, Modelação e Aplicações Computacionais – Centro de Matemática da Universidade do Porto (GEMAC-CMUP), anafjanuario@hotmail.com*

Sónia Gouveia
*Instituto de Engenharia Electrónica e Telemática de Aveiro, Universidade de Aveiro (IEETA-UA) e GEMAC-CMUP*, sonia.gouveia@fc.up.pt

Joaquim Pinto da Costa
*FCUP e CMUP*, jpcosta@fc.up.pt

Maria Isabel Sá
*Serviço de Ginecologia e Obstetrícia, Unidade Maternidade Júlio Dinis- Centro Hospitalar do Porto (MJD-CHP),* misabelrcsa@gmail.com

Alexandra Almeida
*Serviço de Neonatologia, MJD-CHP*, maria.alexandra.almeida@gmail.com

Carmen Carvalho
*Serviço de Neonatologia, MJD-CHP*, carmencarvalho01@gmail.com

Joaquim Saraiva
*Serviço de Ginecologia e Obstetrícia, MJD-CHP*, saraivajp@hotmail.com

Miguel Fonte
*Serviço de Neonatologia, MJD-CHP*, miguelfonte@gmail.com

Paula Soares
*Serviço de Neonatologia, MJD-CHP*, mpccsoares@hotmail.com

## Objectives

Predictive models are useful tools to help clinical decision and help parental counseling, when dealing with extremely premature newborns. In this work, one year mortality of these infants is predicted with classification trees, which are models easy to interpret that provide a clear and logical representation of the data structure.

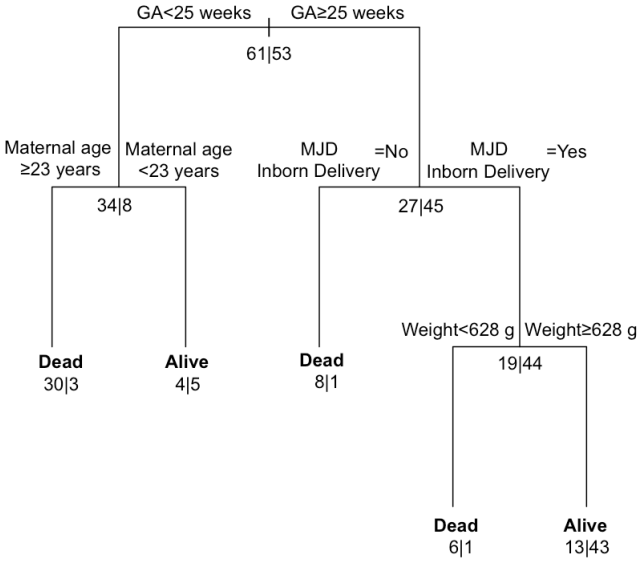
## Methods

Data was collected from 205 newborns (<28 weeks of gestation) followed up at MJD-CHP from 2000 to 2009, comprising one year mortality (yes/no) and 26 variables related to pregnancy and delivery, infant conditions at birth and selected neonatal procedures. There were considered 163 infants, after excluding stillbirths, delivery room death and missing outcomes. The sample was randomly divided into training (70%) and test (30%) sets to construct and validate the predictive model with independent samples, thus avoiding optimistic performance measures. Trees were obtained by recursive partitioning: the root node contains all observations and it is split into 2 nodes and each, by its turn, is divided into 2 other nodes and so on. The variable and cutoff value selected in each node were chosen to obtain a greatest outcome separation. Therefore, the most discriminating variables are typically near the tree top.

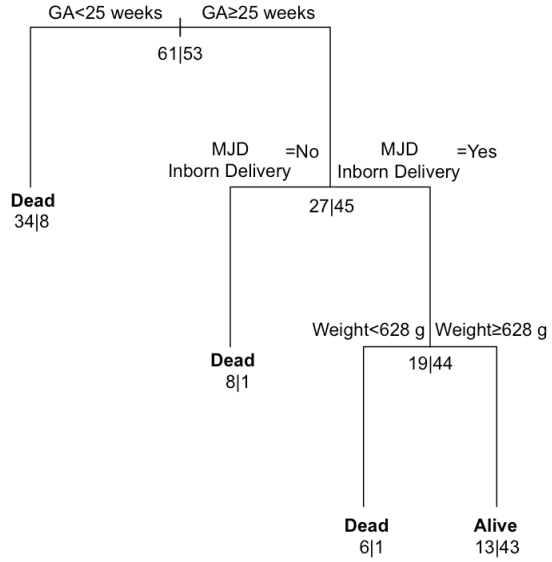
## Results

Figure 1**(a)** presents the classification tree, highlighting GA, MJD Inborn Delivery, Weight and Maternal age as the most important mortality predictors, with predictive accuracy of 61.2%. The tree was then pruned using cross validation, in order to increase predictive efficiency and reduce classification overfitting to the training set. Figure 1**(b)** shows that the pruned tree excluded Maternal age and, as reported in Figure1**(c)**, the overall predictive accuracy increased to 69.4%, with 64.3% and 76.2% of correct predictions in dead/alive outcome. The accuracy of 69.4% for the tree model should be compared with the performance associated with a random prediction, i.e., classify all newborns either as dead or as alive. In the test sample, predictive accuracies for random dead and alive prediction were 57.1% and 42.9%, respectively. The accuracy of the classification tree is higher than that of the random prediction, pointing out the added value of these trees in the prediction of extremely premature newborns.

(a)




(b)




**(c)**

| | | Observed | | Overall Accuracy (%) |
|---|---|---|---|---|
| | | **dead** | **alive** | |
| Predicted | **dead** | 18 | 5 | |
| | **alive** | 10 | 16 | |
| Accuracy (%) | | 64.3% | 76.2% | 69.4% |

**Figure 1:** Classification trees to predict mortality outcome of extreme premature newborns before (a) and after pruning (b): e.g., in (a) a newborn with GA ≥ 25 weeks, Inborn Delivery at MJD and Weight < 628 g is predicted to died. In each node observed (#dead|#alive) on training sample are provided. Predictive performances were estimated on test sample. Acronyms: GA stands for Gestational age.

## Conclusion

The classification tree has important variables for predict mortality: Gestational age, MJD Inborn Delivery and Weight. Our results are agreement with our previous work and literature. The performance of classification tree is higher than that when considering all newborns predicted in the same class.

# Prediction of One Year mortality in extremely premature newborns using classification trees

A.Januário[1,2] S. Gouveia[3,2] J.Pinto da Costa[1,2] M.I.Sá[4] A. Almeida[4] C. Carvalho[4] J.P. Saraiva[4] M. Fonte[4] P. Soares[4]

## Institutions

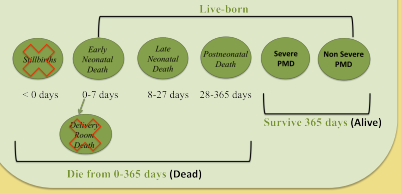
[1] Departamento de Matemática, Faculdade de Ciências, Universidade do Porto
[2] Gabinete de Estatística, Modelação e Aplicações Computacionais, Centro de Matemática da Universidade do Porto (GEMAC - CMUP)
[3] Instituto de Engenharia Electrónica e Telemática de Aveiro, Universidade de Aveiro (IEETA - UA)
[4] Unidade Maternidade de Júlio Dinis – Centro Hospitalar do Porto (MJD)

## Goal

- Predictive models are useful tools to help clinical decision and counseling, when dealing with extremely premature newborns.
- Classification trees are predictive models that provide a clear and logical representation of the data structure.
- **Goal:** To predict one year mortality and two years morbidity of extremely premature newborns (< 28 weeks of gestational age - GA)

## Data

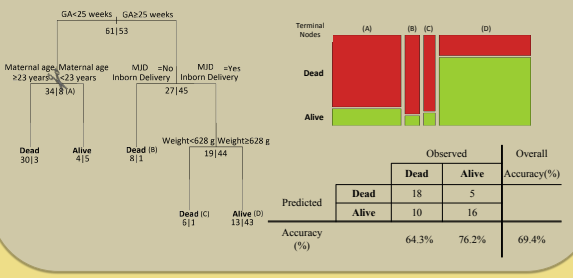
- **205 newborns** followed up at **MJD** (2000 – 2009).
- **163 infants**, after excluding stillbirths, delivery room death and missing outcomes.
- **26 variables** relating to pregnancy and delivery, infant conditions at birth and selected neonatal procedures.
- **2 outcomes:** one year mortality (Dead/Alive) and two years morbidity (Severe/Non Severe) assessed by PMD (Psychomotor development).




## Results

### Mortality

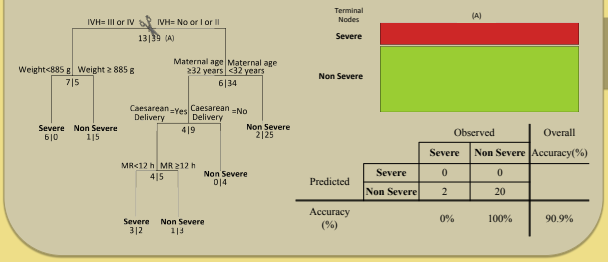
Completed and Pruned Tree (🔧) — Performance Classification




| Predicted \ Observed | Dead | Alive | Overall Accuracy(%) |
|---|---|---|---|
| Dead | 18 | 5 | |
| Alive | 10 | 16 | |
| Accuracy (%) | 64.3% | 76.2% | 69.4% |

### Morbidity

Completed and Pruned Tree (🔧) — Performance Classification



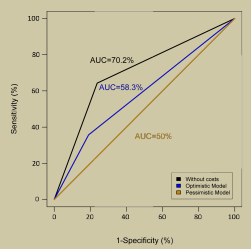

| Predicted \ Observed | Severe | Non Severe | Overall Accuracy(%) |
|---|---|---|---|
| Severe | 0 | 0 | |
| Non Severe | 2 | 20 | |
| Accuracy (%) | 0% | 100% | 90.9% |

## With costs(*)

### Mortality

| Predicted \ Observed | Dead | Alive | Overall Accuracy(%) |
|---|---|---|---|
| Dead | 10 ; 28 | 4 ; 21 | |
| Alive | 18 ; 0 | 17 ; 0 | |
| Accuracy (%) | 35.7% ; 100% | 80.95% ; 0% | 55.1% ; 57.1% |

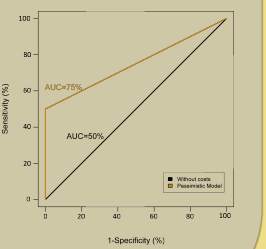
ROC CURVE




### Morbidity

| Predicted \ Observed | Severe | Non Severe | Overall Accuracy(%) |
|---|---|---|---|
| Severe | 1 | 0 | |
| Non Severe | 1 | 20 | |
| Accuracy (%) | 50% | 100% | 95.5% |

ROC CURVE




(*) In both cases 2 cost was applied because we found that it is sufficient to improve the classification of positive/negative outcome

## Methods

- The sample was randomly divided into **training** (70%) and **test** (30%)

**Tree construction in training set, [2]**

- **[Completed Tree]** was obtained by binary recursive partitioning, where in each node, the variable and cutoff selected were chosen to obtain the best partition
- **[Pruned tree]** was obtained after excluding the less important variables of the completed tree

**Tree validation in test set**

- Independent set, thus avoiding optimistic performance measures

**3 approaches**

- **Without costs**
- With costs: **Optimistic** [increase correct predictions of the positive outcome (Alive / Non Severe)]
- With costs: **Pessimistic** [increase correct predictions of the negative outcome (Dead / Severe)]

## Conclusions

- Mortality classification trees showed acceptable discrimination (overall accuracy of 69.4% and AUC of 70.2%), [3]:
  - negative mortality outcome for newborns with GA<25 weeks [1];
  - positive mortality outcome for newborns with GA≥25 weeks, born at MJD and with Weight≥628 g.
- It was not possible to obtain a Morbidity classification tree, probably due to uneven class frequencies:
  - All newborns are classified as with non severe PMD.
- When introducing costs, pessimistic tree on mortality was not significant (overall accuracy of 57.1% and AUC of 50.0%), whereas on morbidity showed accuracy of 95.5% and discrimination of 75% (AUC)
- The majority of variables selected by this method is in accordance with our previous work [4].

## References

[1] Ambalavanan, N., Baibergenova, A., Carlo, W.A., Saigal, S., Schmidt, B., Thorpe, K.E., and TIPP Investigators (2006). Early Prediction of Poor Outcome In Extremely Low Birth Weight Infants By Classification Tree Analysis. *Journal of Pediatrics*, 148, 348 – 44.
[2] Breiman, L., Friedmann, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Florida: Chapman & Hall.
[3] Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*, John Wiley and Sons.
[4] Januário, A., Gouveia, S., Pinto da Costa, J., Sa, M.I., Almeida, A., Carvalho,C., Saraiva, J.P., Fonte, M., Soares, P. (2012). *Logistic Regression In the Study of One-Year Mortality in Portuguese Extremely Premature Newborns: Impact of Protocol and Intrinsic Risk Factors*. Abstract JOCLAD'12 , 81-84.

Ref[4]. Apresentação oral em XX Congresso da SPE

**Prediction methods for mortality and morbidity prognostic of Portuguese extremely premature newborns**

Ana Januário
*Departamento de Matemática, Faculdade de Ciências da Universidade do Porto (FCUP) e Gabinete de Estatística, Modelação e Aplicações Computacionais - Centro de Matemática da Universidade do Porto (GEMAC-CMUP), anafjanuario@hotmail.com*

Sónia Gouveia
*Instituto de Engenharia Electrónica e Telemática de Aveiro, Universidade de Aveiro (IEETA-UA) e GEMAC-CMUP, sonia.gouveia@fc.up.pt*

Joaquim Pinto da Costa
*FCUP e CMUP, jpcosta@fc.up.pt*

Maria Isabel Sá
*Serviço de Ginecologia e Obstetrícia, Unidade Maternidade Júlio Dinis - Centro Hospitalar do Porto (MJD-CHP), misabelrcsa@gmail.com*

Alexandra Almeida
*Serviço de Neonatologia, MJD-CHP, maria.alexandra.almeida@gmail.com*

Carmen Carvalho
*Serviço de Neonatologia, MJD-CHP, carmencarvalho01@gmail.com*

Joaquim Saraiva
*Serviço de Ginecologia e Obstetrícia, MJD-CHP, saraivajp@hotmail.com*

Miguel Fonte
*Serviço de Neonatologia, MJD-CHP e Transporte Inter - Hospitalar Pediátrico do Norte, Centro Hospitalar São João, miguelfonte@gmail.com*

Paula Soares
*Serviço de Neonatologia, MJD-CHP, mpccsoares@hotmail.com*

**Keywords**: discriminant analysis, extremely premature newborns

**Abstract**: In recent years, the mortality rate of extremely premature newborns (<28 weeks of gestation) has decreased due to the advances in perinatal care. However, when surviving, the risk for late disabilities(morbidity) and health complications is still high [5]. The purpose of this study is to construct predictive models based on discriminant analysis, to predict one year mortality and two years morbidity of a Portuguese population of extremely premature newborns.

**Preliminary results**

Data were collected from 205 extremely premature newborns followed up at MJD-CHP (2000 - 2009) with approval of the CHP Ethics Commitee. The dataset comprised two infant outcomes, one-year mortality (yes/no) and two years morbidity (yes/no), plus 26 variables related to pregnancy and delivery, infant conditions at birth and selected neonatal procedures. In our previous work, we studied 17 of the 26 variables to determine earlier mortality risk factors by simple logistic regression [2, 3]. The joint effect of these risk factors was assessed by multivariate logistic regression and results showed that only 30% of mortality variance was explained by this model. Therefore, in this work we explore all 26 available variables to predict mortality/morbidity. From the 205 initial observations we considered 163 infants, after excluding stillbirths, delivery room death and missing outcomes. Moreover, to properly evaluate the constructed models the sample was randomly divided into training (70%) and test (30%) samples. The final training sample size of $n$=104 restricts the number of variables in the logistic model and, following a recommendation of 10 events per variable [4], only up to 5 variables should be considered. Therefore, the original set of 26 variables had to be reduced, either choosing one or merging variables from a subset of highly associated variables. From the 3 continuous variables, weight and gestational age were considered as they were previously identified as significant mortality risks factors [2, 3, 5]. The pairwise associations between the 23 categorical variables were explored by Cramer's coefficient V = $X^2/k$, where $X^2$ is the chi-square statistics, $k = n[min(r, c) - 1]$, $n$ is the sample size and ($r$,$c$) are the number of (rows, columns) of contingency table. This coefficient ranges from 0 to 1, with V=0 indicating no association between two variables. Hierarchical clustering with

dissimilarity matrix 1-**V** was used to find subsets of associated variables. Different types of linkage (single-link, complete-link, average-link, Ward's method, center of gravity) were compared by delta deformation $\delta$ and cophenetic correlation $\gamma$. Figure 1 presents the dendrogram by group average clustering, which shown to have simultaneously smallest $\delta$ and largest $\gamma$. The dendrogram shows that most of the variables are grouped with high linkage values, and any considered subsets would include quite heterogeneous variables.

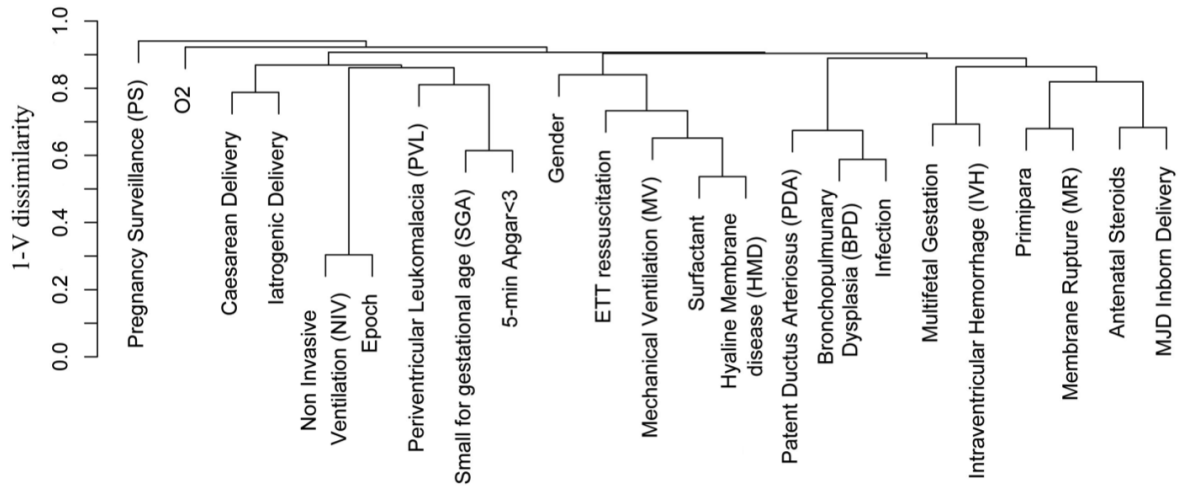



Figura 1: Dendrogram by group average clustering based on 1-**V** dissimilarity. Most associated variables are NIV and Epoch, whereas PS is the last variable included in the linkage.

An alternative approach to obtain the optimal logistic regression model up to 5 variables is to consider the Furnival and Wilson algorithm [1]. In this work, the logistic regression mortality/morbidity models will be compared with classifications trees and neuronal networks through ROC curves [1].

**Acknowledgments:** This work was supported by FEDER through COMPETE programme and FCT (proj. FCOMP-01-0124-FEDER-022682, IEETA, www.ieeta.pt and FCOMP-01-0124-FEDER-022656, CMUP, www.fc.up.pt/cm
The authors also thank funding from Abbot Laboratórios.

# Referências

[1] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition, Springer.

[2] Januário, A., Gouveia, S., Pinto da Costa, J., Sá, M.I., Almeida, A., Carvalho,C., Saraiva, J.P., Fonte, M., Soares, P. (2012). Logistic Regression In the Study of One-Year Mortality in Portuguese Extremely Premature Newborns: Impact of Protocol and Intrinsic Risk Factors. *Abstract JOCLAD'12*, 81–84.

[3] ——— (2012). Risk factors Associated with One-Year Mortality of Extremely Premature Newborns in Portugal. *Abstract IJUP'12*, 118.

[4] Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R. (1996). A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *J Clin Epidemiol* 49(12), 1373–1379.

[5] The Express Group Members (2009). One-year Survival of Extremely Preterm Infants After Active Perinatal Care in Sweden. *JAMA* 301(21), 2225–2233.

Ref [5]. Resumo da apresentação oral em XLI Jornadas Nacionais de Neonatologia da Sociedade Portuguesa de Pediatria

**Management and Outcome of Extremely Premature Infants**

Maria Isabel Sá
*Serviço de Ginecologia e Obstetrícia, Unidade Maternidade Júlio Dinis, Centro Hospitalar do Porto, Porto, misabelrcsa@gmail.com*
Miguel Fonte
*Serviço de Neonatologia, Unidade Maternidade Júlio Dinis, Centro Hospitalar do Porto, miguelfonte@gmail.com*
Joaquim Saraiva
*Serviço de Ginecologia e Obstetrícia, Unidade Maternidade Júlio Dinis, Centro Hospitalar do Porto, Porto, saraivajp@hotmail.com*
Cármen Carvalho
*Serviço de Neonatologia, Unidade Maternidade Júlio Dinis, Centro Hospitalar do Porto, Porto, carmencarvalho01@gmail.com*
Paula Soares
*Serviço de Neonatologia, Unidade Maternidade Júlio Dinis, Centro Hospitalar do Porto, Porto, mpccsoares@hotmail.com*
Ana Januário
*Departamento de Matemática, Faculdade de Ciências da Universidade do Porto e Gabinete de Estatística, Modelação e Aplicações Computacionais – Centro de Matemática da Universidade do Porto, Porto, anafjanuario@hotmail.com*
Sónia Gouveia
*Instituto de Engenharia Electrónica e Telemática de Aveiro e Gabinete de Estatística, Modelação e Aplicações Computacionais – Centro de Matemática da Universidade do Porto, Porto, sonia.gouveia@fc.up.pt*
Alexandra Almeida
*Serviço de Neonatologia, Unidade Maternidade Júlio Dinis, Centro Hospitalar do Porto, Porto, maria.alexandra.almeida@gmail.com*

Objective
In the last decades, survival of extremely premature infants has improved but there is still significant morbidity among this group of children. We intend to describe medical management in the perinatal-neonatal period and the outcome at 18-24 months of a group of extremely premature infants. Our goal is to evaluate whether specific attitudes or characteristics explain higher rates of survival and survival without long-term severe disabilities.

Methods
A retrospective study was conducted including all 205 liveborn and stillborn infants with gestational age between $22w^{0d}$ and $26w^{6d}$, born in the Obstetric Unit or transferred to the Neonatology Unit of our institution from January 2000 to December 2009. We collected variables related to management in the prenatal and neonatal period, infant performances and psychomotor development at 18-24 months. Significant associations between variables and outcomes were identified by chi-square test or t-test, and multivariate logistic models were used to describe and predict survival/morbidity.

Results
Gestational age, antenatal corticotherapy, cesarean section, inborn delivery and weight were associated with an increasing survival ($p<0.05$), while absence of intraventricular hemorrhage (IVH) grade 3-4 and periventricular leukomalacia was associated with survival without severe neurossensorial deficit ($p<0.05$). According to the multivariate models, advanced gestational age (OR=0,353), increasing weight (OR=0,996) and antenatal corticotherapy (OR=0,150) were associated with survival increase. IVH was associated with morbility increase (OR between 9,006 and 13,294). These models predicted correctly survival in 78,1% and severe morbidity in 87,8% of the cases.

Conclusions
Our results are consistent with other published works and the survival/morbidity models might be a valuable tool providing some insight in the prediction of the outcome of the extreme premature neonates, guide decision making and help parental counseling.

Key words: Extreme premature, Outcome, Psychomotor development, Survival

# Management and Outcome of Extremely Premature Infants

Maria Isabel Sá[1], Miguel Fonte[2], Joaquim Saraiva[1], Cármen Carvalho[2], Paula Soares[2], Ana Januário[3], Sónia Gouveia[3], Alexandra Almeida[2]
[1]*Serviço de Ginecologia e Obstetrícia, Unidade Maternidade Júlio Dinis, Centro Hospitalar do Porto*
[2]*Serviço de Neonatologia, Unidade Maternidade Júlio Dinis, Centro Hospitalar do Porto*
[3]*Departamento de Matemática, Faculdade de Ciências da Universidade do Porto e Gabinete de Estatística, Modelação e Aplicações Computacionais – Centro de Matemática da Universidade do Porto*

## Introduction

In the last decades, survival of extremely premature infants has improved but there is still significant morbidity among this group of children. We intend to describe medical management in the perinatal-neonatal period and the outcome at 18-24 months of a group of extremely premature infants. Our goal is to evaluate whether specific attitudes or characteristics explain higher rates of survival and survival without long-term severe disabilities (cerebral palsy, neurossensorial blindness and neurossensorial deafness with need of auditive prosthesis).

## Methods

A retrospective study was conducted including all 205 liveborn and stillborn infants with gestational age between 22w$^{0d}$ and 26w$^{6d}$, born in the Obstetric Unit or transferred to the Neonatology Unit of our institution from January 2000 to December 2009. We collected variables related to management in the prenatal and neonatal period, infant performances and psychomotor development at 18-24 months. Significant associations between variables and outcomes were identified by chi-square test or t-test, and multivariate logistic models were used to describe and predict survival/morbidity.

## Results

**Total sample 205** = **Stillbirth 30** (14,6%) + **Delivery Room Death 6** (2,9%) + **Early Neonatal Death 67** (32,7%) + **Death before 18-24 months 22**(10,7%) + **Alive With** Major disabilities **15** (7,3%) + **Alive Without** Major disabilities **59** (28,8%)

1. Sample´s distribution by outcome

| | Total |
|---|---|
| Pregnancy surveillance | 92,1% |
| Multifetal gestation | 32,2% |
| Antenatal steroids | 73,2% |
| Iatrogenic delivery | 9,9% |
| Vaginal delivery | 48,5% |
| Male gender | 57,6% |

2. Total sample caracterization

| | Total |
|---|---|
| Ressuscitation with ETT | 69,7% |
| 5´Apgar ≤3 | 11,2% |
| Surfactant | 86,3% |
| One year survival | 45,7% |

3. Liveborn caracterization

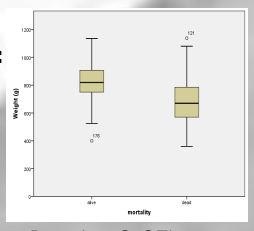
| Morbidity | Infants survival > 24h | Infants survival > 1 year |
|---|---|---|
| Intraventricular hemorrhage III-IV | 30,1% | 16,3% |
| Periventricular leukomalacia | 3,7% | 2,5% |
| Hyaline membrane disease | 90,2% | 91,3% |
| Bronchopulmonary dysplasia | 36,8% | 77,5% |
| Persistent Ductus Arteriosus | 38,7% | 56,3% |
| Infection | 55,2% | 78,8% |

4. Morbidity in Infants admitted to the Intensive Care Unit (all vs. Infants alive at the age of follow-up)

## Mortality and morbility models

### a) Univariate analysis

**Increase in survival associated to (p<0.05):**
-Gestational Age
-Antenatal Steroids
-Cesarean Section
-Inborn delivery
-Weight

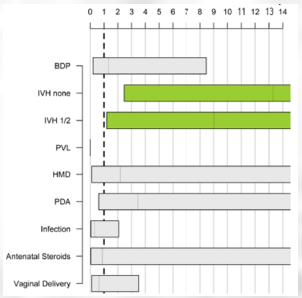



**Survival without major disabilities associated to (p<0.05):**
-Absence of intraventricular hemorrhage III-IV
-Absence of periventricular leukomalacia

### b) Multivariate logistic regression analysis

**Mortality odds ratio (OR) and 95% CI**
**OR<1 indicates decreased risk of mortality**


**Morbidity odds ratio (OR) and 95% CI**
**OR>1 indicates decreased risk of morbidity**




### b) Multivariate logistic regression models – Predicting Outcome

1. Hosmer-Lemeshow test for the goodness of fit of the multivariate models indicate both models describe adequately the data
**p=0.194 for mortality and p=0.289 for morbidity**
2. The large Nagelkerke's R2 values indicated that the models explain a substantial amount of proportion of mortality/morbidity variance
**R2=52,6% for mortality and R2=42,4% for morbidity**
3. Therefore, these multivariate models are expected to predict adequately mortality/morbidity in extremely premature newborns
**Predicting capacity – 78,1% for mortality / Predicting capacity – 87,8% for morbidity**

## Conclusions

Our results are consistent with other published works and the survival/morbidity models might be a valuable tool providing some insight in the prediction of the outcome of the extreme premature neonates, guide decision making and help parental counseling.

References
Berger,T et al; Perinatal care at the limit of viability between 22 and 26 completed weeks of gestation in Switzerland; The European Journal of Medical Sciences; 141, w13280; 2011; Marlow, N et al; Neurologic and Developmental Disability at Six Years of Age after Extremely Preterm Birth; The New England Journal of Medicine; 352, 1, 9-18; 2005; Seri, I et al; Limits of viability: definition of the gray zone; Journal of Perinatology, 28, 84-88; 2008; Kaiser, J et al; Hospital Survival of Very-Low-Birth-Weight Neonates from 1977 to 2000; Journal of Perinatology; 24, 343-350; 2004; The EXPRESS Group; One-Year Survival of Extremely Preterm Infants After Active Perinatal Care in Sweden; Journal of American Medical Association; 301(21):2225-2233; 2009; Field, D et al; Survival of extremely premature babies in a geographically defined population: prospective cohort study of 1994-9 compared with 2000-5; British Medical Journal; online edition; 2008; Wang, Y et al; Effect of pregestational maternal, obstetric and perinatal factors on neonatal outcome in extreme prematurity; Arch Gynecol Obstet, 404-011, 1870-5; 2011; Bodeau-Livinec, F et al; Impact of Intensive Care Practices on Short-Term and Long-term Outcomes for Extremely Preterm Infants: Comparison Between the British Isles and France; Pediatrics; 122, 5, e1014-e1021; 2008; Blanco, F et al; Ensuring Accurate Knowledge of Prematurity Outcomes for Prenatal Counseling; Pediatrics; 115, 4, e478-e487; 2005; Lorenz, J et al; A Quantitative Review of Mortality and Developmental Disability in Extremely Premature Newborns; Arch Pediatr Adolesc Med; 152, 425-435; 1998; Vento, M etal; The First Golden Minutes of the Extremely-Low-Gestational-Age Neonate: A Gentle Approach; Neonatology; 95, 286-298; 2009; Kadri, H et al; The incidence, timing, and predisposing factors of germinal matrix and intraventricular hemorrhage (GMH/IVH) in preterm neonates; Childs Nerv Syst; 22, 1086-1090; 2006; Westra, S et al; Reader Variability in the use of Diagnostic Terms to describe White Matter Lesions Seen on Cranial Scans of Severely Premature Infants: The ELGAN Study; J Clin Ultrasound; 38,409-419; 2010; Kamoji, V et al; Extremely Growth-Retarded Infants: Is There a Viability Centile?; Pediatrics; 118, 2, 758-763; 2006;

misabelrcsa@gmail.com

Ref [6]. Resumo da apresentação oral em COGI

**Management and Outcome of Extremely Premature Infants**

Maria Isabel Sá[1], Miguel Fonte[2], Cármen Carvalho[2], Paula Soares[2], Alexandra Almeida[2], Ana Januário[3], Sónia Gouveia[3,4], Joaquim Saraiva[1]

[1]*Serviço de Ginecologia e Obstetrícia, Unidade Maternidade Júlio Dinis, Centro Hospitalar do Porto, Porto*
[2]*Serviço de Neonatologia, Unidade Maternidade Júlio Dinis, Centro Hospitalar do Porto*
[3]*Departamento de Matemática, Faculdade de Ciências da Universidade do Porto e Gabinete de Estatística, Modelação e Aplicações Computacionais – Centro de Matemática da Universidade do Porto, Porto*
[4]*Instituto de Engenharia Electrónica e Telemática de Aveiro, Aveiro*

Problem Statement
Over last decades, survival of extremely premature infants improved but there´s still significant morbidity among this group. We intend to describe management in the perinatal-neonatal period and outcome at 18-24 months of a group of extremely premature infants. Our goal is to evaluate whether specific attitudes/characteristics explain higher survival and survival without severe disabilities rates.

Methods
We conducted a retrospective study including 205 liveborn/stillborn infants (gestational age $22w^{0d}$ to $26w^{6d}$), born in the Obstetric Unit or transferred to the Neonatology Unit from January 2000 to December 2009. We collected variables related to management in the prenatal/neonatal period, infant performances and psychomotor development at 18-24 months. Significant associations between variables and outcomes were identified by chi-square test or t-test, and multivariate logistic regression models were used to describe and predict mortality/morbidity.

Results
Advanced gestational age, antenatal corticotherapy, cesarean section, inborn delivery and increased weight were associated with survival($p<0.05$), while absence of intraventricular hemorrhage(IVH) grade 3-4 and periventricular leukomalacia was associated with survival without severe neurossensorial deficit($p<0.05$). According to multivariate models, advanced gestational age(OR=0,353), increased weight(OR=0,996) and antenatal corticotherapy(OR=0,150) were associated with lower mortality risk. IVH grade 3-4 was associated with higher morbility risk(OR=16,931). Statistical analysis indicated that both models describe adequately the data and predict correctly mortality and severe morbidity in 78,1% and 85,7% of the cases, respectively.

Conclusions
Our results are consistent with literature and mortality/morbidity models might be valuable tools providing insight in the prediction of the outcome of the extreme premature neonates and help parental counseling.

# Referências


Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Willey& Sons, INC.

Ambalavanan, N., Baibergenova, A., Carlo, W., Saigal, S., Schmidt, B., and Thorpe, K. (2006). Early prediction of poor outcome in extremely low birth weight infants by classification tree analysis. *Journal of Pediatrics*, 148:438–444.

Antunes, L., Bento, M. J., and Mendonça, D. (2011). Imputação múltipla - uma aplicação ao tratamento de dados omissos em análise de sobrevivência de doentes oncológicos. In *Book of Abstracts of XIV Congresso da Sociedade Portuguesa de Estatística*, pages 271–272.

Arad, I., Braunstein, R., and Oz, B. B. (2008). Neonatal outcome of inborn and outborn extremely low birth weight infants: Relevance of perinatal factors. *Israel Medical Association Journal (IMAJ)*, 10:457–461.

Araújo, B. F., Zatti, H., Filho, P. F. O., Coelho, M. B., Olmi, F. B., Guaresi, T. B., and Madi, J. M. (2011). Effect of place of birth and transport on morbidity and mortality of preterm newborns. *J. Pediatr (Rio J.)*, 87(3):257–262.

Austin, P. C., V.Tu, J., and Lee, D. S. (2010). Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *Journal of Clinical Epidemiology*, 63(10):1145–1155.

Bacak, S. J., Baptiste-Roberts, K., Amon, E., Ireland, B., and Leet, T. (2005). Risk factors for neonatal mortality among extremely-low-birth-weight infants. *American Journal of Obstetrics and Gynecology*, 192:862–867.

Bagley, S. C., White, H., and Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*, 54(10):979–985.

Batista, G. E., C.Prati, R., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6(1):20–29.

Beckaman, C. R., Ling, F. W., Barzansky, B. M., Herbert, W. N., and Laube, D. W. (2010). *Obstetrics and Gynecology*, chapter 17- Multifetal Gestation. Lippincott Williams & Wilkins, a Wolters Kluwer business, 6 edition.

Bellazzi, R. and Zupan, B. (2006). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77(2):81–97.

Berger, T. M., Bernet, V., Alama, S. E., Faucère, J.-C., Hosli, I., Irion, O., Kind, C., Latal, B., Nelle, M., Pfister, R. E., Surbek, D., Truttmann, A. C., Wisser, J., and Zimmermann, R. (2011). Perinatal care at the limit of viability between 22 and 26 completed weeks of gestation in switzerland. *The European Journal of Medical Sciences*, 141:1–13.

Bhaumik, U., Aitken, I., Kawachi, I., Orav, J., and Lieberman, E. (2004). Narrowing of sex differences in infant mortality in massachusetts. *Journal of Perinatology*, 24:94–99.

Blumenfeld, Y. J., Lee, H. C., Goud, J. B., Langen, E. S., Jafari, A., and El-Sayed, Y. Y. (2010). The effect of preterm premature rupture of membrane on neonatal mortality rates. *Obstetrics & Gynecology*, 116(6):1381–1386.

Boussicault, G., Branger, B., Savagner, C., and Rozé, J. (2012). Survie et devenir neurologique à l'âge corrigé de 2 ans des enfants nés extrêmement prématurés. *Archieves de Pédiatrie*, (19):381–390.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification And Regression Trees*. Chapman & Hall.

Camdeviren, H. A., Yazici, A. C., Akkus, Z., Bugdayci, R., and Sungur, M. A. (2007). Comparison of logistic regression model and classification tree: An application to postpartum depression data. *Expert Systems with Applications*, 32(987-994).

Carlo, W. A., McDonald, S. A., Fanaroff, A. A., Vohr, B. R., Stoll, B. J., Ehrenkranz, R. A., Andrews, W. W., Wallace, D., Das, A., Bell, E. F., Walsh, M. C., Laptook, A. R., Shankaran, S., Poindexter, B. B., C.Hale, E., Newman, N. S., Davis, A. S., Schibler, K., Kennedy, K. A., Sánchez, P. J., Meurs, K. P. V., Goldberg, R. M., Watterberg, K. L., Faix, R. G., III, I. D. F., and Higgins, R. D. (2011). Association of antenatal corticosteroids with mortality and neurodevelopmental outcomes among infants born ate 22 to 25 weeks' gestation. *JAMA*, 306(21):2348–2358.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Duxbury Press.

Castro, M. P., Moura, M. D. R., de Souza Rugolo, L. M. S., and Margotto, P. R. (2011). Limite de viabilidade no moderno cuidado intensivo neonatal - análise além da idade gestacional. *Comunicação em Ciências da Saúde*, 22(1):101–112.

Chatterjee, S., Hadi, A. S., and Price, B. (2000). *Regression Analysis By Example*. JOHN WILEY & SONS, INC, USA, third edition.

Colombet, I., Ruelland, A., Chatellier, G., Gueyffier, F., Degoulet, P., and Jaulent, M.-C. (2000). Models to predict cardiovascular risk:comparison of cart, multilayer perceptron and logistic regression. In *Proceedings AMIA Symposium*, pages 156–160.

da Graça, L. M. (2010). *Medicina Materno-Fetal*. Lidel, 4ª edition.

Dobson, A. J. (2002). *An Introduction To Genelarized Linear Models*. CHAPMAN & HALL/CRC, USA, second edition.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons, INC, 2 edition.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate statistical Modelling Based on Generalized Linear Models*. Springer, New York, second edition.

Faraway, J. J. (2006). *Extending the Linear Model with R*. CHAPMAN & HALL/CRC.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874.

Fonseca, J. M. M. R. (1994). Indução de Árvores de decisão: Hitclass - proposta de um algoritmo não paramétrico. Master's thesis, Departamento de Informática, Faculdade de Ciências e Tecnologia, Univsersidade Nova de Lisboa.

Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *TECHNOMETRICS*, 16(4).

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764.

Gouveia, S. (2009). *Contributions to the analysis of short-term cardiovascular coupling*. PhD thesis, Faculdade de Ciências da Universidade do Porto.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, pages 1157–1182.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, 2 edition.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons, INC, second edition.

Hu, B., SHao, J., and Palta, M. (2006). Pseudo $r^2$ in logistic regression model. *Statistica Sinica*, pages 847–860.

Januário, A., Gouveia, S., da Costa, J. P., Sá, M. I., Almeida, A., Carvalho, C., Saraiva, J. P., Fonte, M., and Soares, P. (2012a). Logistic regression in the study of one-year mortality in portuguese extremely premature newborns: impact of protocol and intrinsic risk factors. In *Book of Abstracts of XIX Jornadas de Classificação e Análise de Dados (JOCLAD2012)*, pages 81–84. Associação Portuguesa de Classificação e Análise de Dados (CLAD)/ Instituto Politécnico de Tomar.

Januário, A., Gouveia, S., da Costa, J. P., Sá, M. I., Almeida, A., Carvalho, C., Saraiva, J. P., Fonte, M., and Soares, P. (2012b). Prediction methods for mortality and morbidity prognostic of portuguese extremely premature newborns. In *Book of Abstracts of XX Congresso da Sociedade Portuguesa de Estatística*, pages 269–273. Sociedade Portuguesa de Estatística (SPE).

Januário, A., Gouveia, S., da Costa, J. P., Sá, M. I., Almeida, A., Carvalho, C., Saraiva, J. P., Fonte, M., and Soares, P. (2012c). Prediction of one year mortality in extremely premature newborns using classification trees. In *Book of Abstracts of $4^a$as Jornadas de Iniciação à Investigação Clínica*, pages 125–126. Departamento de Ensino, Formação e Investigação do Centro Hospitalar do Porto (DEFI-CHP).

Januário, A., Gouveia, S., da Costa, J. P., Sá, M. I., Almeida, A., Carvalho, C., Saraiva, J. P., Fonte, M., and Soares, P. (2012d). Risk factors associated with one-year mortality of extremely premature newborns in portugal. In *Book of Abstracts of 5th Meeting of Young Researchers of University of Porto (IJUP'12)*, page 118. Universidade do Porto.

Kaempf, J. W., Tomlinson, M. W., Campbell, B., Ferguson, L., and Stewart, V. T. (2009). Counseling pregnant woman who may deliver extremely premature infants: Medical care guidelines, family choices, and neonatal outcomes. *Journal of the American Academy of Pediatrics*, 123(6):1509–1515.

Kaiser, J. R., Tilford, J. M., SImpson, P. M., Salhab, W. A., and Rosenfeld, C. R. (2004). Hospital survival of very-low-birth-weight neonates from 1977 to 2000. *Journal of Perinatology*, 24:343–350.

Kitsantas, P., Hollander, M., and Li, L. (2006). Using classification trees to assess low birth weight outcomes. *Artificial Intelligence in Medicine*, 38(3):275–289.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, pages 1137–1143. Morgan Kaufmann.

Kohavi, R. and Quinlan, R. (1999). Decision tree discovery. In *Handbook of Data Mining and Knowledge Discovery*, pages 267–276. University Press.

Lantos, J. D. and Meadow, W. (2009). Variation in the treatment of infants born at the borderline of viability. *Journal of the American Academy of Pediatrics*, 123(6):1588–1590.

Lasko, T. A., Bhagwat, J. G., Zou, K. H., and Ohno-Machado, L. (2005). The use of receiver operating characteristics curves in biomedical informatics. *Journal of Biomedical Informatics*, 38:404–415.

Liebetrau, A. (1983). *Measures of association*. Sage University Paper.

Lindsey, J. K. (1997). *Applying Generalized Linear Models*. Springer.

McCrea, H. J. and Ment, L. R. (2008). The diagnosis, management and postnatal prevention of intraventricular hemorrhage in the preterm neonate. *Clin Perinatol*, 35(4):777–792.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman& Hall, second edition.

McLeod, A. I. and Xu, C. (2011). *bestglm: Best Subset GLM*. R package version 0.33, disponível em http://CRAN.R-project.org/package=bestglm.

Medlock, S., Ravelli, A. C. J., Tamminga, P., Mol, B. W. M., and Abu-Hanna, A. (2011). Prediction of mortality in very premature infants: A systematic review of prediction models. *PLoS One*, 6(9):e23441.

Mikhno, A. and Ennett, C. M. (2012). Prediction of extubation failure for neonates with respiratory distress syndrome using the mimic-ii clinical database. In *Book of Abstracts of 34th Annual International Conference of the IEEE EMBS*, pages 5094–5097.

Morgan, J. A. and Tatar, J. F. (1972). Calculation of the residual sum of squares for all possible regressions. *Technometrics*, 14(2):317–325.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58:415–434.

Muris, C., Girard, B., Creveuil, C., Durin, L., and Dreyfus, M. (2007). Management of premature rupture od membranes before 25 weeks. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 131:163–168.

Nelder, J. A. and Wedderburn, R. R. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135(3):370–384.

Obuchowskl, N. A. (2003). Receiver operating characteristic curves and their use in radiolody. *Radiology*, 229(1):3–8.

Orlandi, S., Manfredi, C., Bocchi, L., and Scattoni, M. L. (2012). Automatic newborn cry analysis: a non-invasive tool to help autism earky diagnosis. In *Book of Abstracts of 34th Annual International Conference of the IEEE EMBS*, pages 2953–2956.

Peacock, J. L., Marston, L., Marlow, N., Calvert, S. A., and Greenough, A. (2012). Neonatal and infant outcome in boys and girls born very prematurely. *International Pediatric Research Foundation, Inc*, pages 1–6.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemol*, 49(12):1373–1379.

Peixoto, J., Branco, M., Freitas, A., and Dias, C. (2004). Viabilidade. In *Secção de Neonatologia da Sociedade Portuguesa de Pediatria-Consensos nacionais em neonatologia*, pages 11–16.

Peralta-Carcelen, M., Moses, M., Adams-Chapman, I., Gantz, M., and Vohr, B. R. (2009). Stability of neuromotor outcomes at 18 and 30 months of age after extremely low birth weight status. *Pediatrics*, 123(5):e887–e895.

Podgorelec, V., Kokol, P., Stiglic, B., and Rozman, I. (2002). Decision trees: an overview and their use in medicine. *Journal of Medical Systems*, 26(5):445–463.

Prati, R., Batista, G., and Monard, M. (2008). Curvas roc para avaliação de classificadores. *IEE Latin America Transactions*, 6(2):215–222.

Precup, D., Robles-Rubio, C. A., Brown, K. A., L.Kanbar, Kaczmarek, J., Chawla, S., Sant'Anna, G. M., and Kearney, R. E. (2012). Prediction of extubation readiness in extreme preterm infants based on measures of cardiorespiratory variability. In *Book of Abstracts of 34th Annual International Conference of the IEEE EMBS*, pages 5630–5633.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, (1):81–106.

Quinlan, J. (1993). *Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, retrieved from `http://www.R-project.org/`.

Rennie, J. M. (1996). Perinatal management at the lower margin of viability. *Archives of Disease in Childhood*, 74:F214–F218.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rodríguez, G. (2007). Lecture notes on generalized linear models. Retrieved October 15, 2011, from `http://data.princeton.edu/wws509/notes/`.

Sankaran, K., Chien, L.-Y., Walker, R., Seshia, M., Ohlsson, A., Lee, S. K., and the Canadian Neonatal Network (2002). Variations in mortality rates among canadian neonatal intensive care units. *Canadian Medical Association*, 166(2):173–178.

Seri, I. and Evans, J. (2008). Limits of viability:definition of the gray zone. *Journal of Perinatology*, (28):S4–S8.

Shtatland, E. S., Kleinman, K., Cain, E. M., School, H. M., and and, H. P. H. C. (2002). One more time about $r^2$ measures of fit in logistic regression. In *Boof of Abstracts of NESUG 15*. NorthEast SAS Users Group.

Silva, E. B. (2008). Tomada de decisões éticas em prematuros: experiências de médicos e enfermeiras. *Revista Portuguesa de Bioética*, (5):191–206.

Silva, E. B. and Carvalho, A. S. (2008). Questões éticas da prematuridade. In *Bioética e vulnerabilidade*. Almedina.

Smith, P. B., Ambalavanan, N., Li, L., Cotten, C. M., Laughon, M., C.Walsh, M., Das, A., Bell, E. F., Carlo, W., Stoll, B. J., Shankaran, S., and Laptook, A. R. (2012). Approach to infants born at 22 to 24 week's gestation: Relationship to outcomes of more-mature infants. *Journal of American Academy of Pediatrics*, 129(6):e1508–e1516.

Soibelman, L. and Kim, H. (2002). Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering*, 16(1):39–48.

Stevenson, D. K., Verter, J., Fanaroff, A. A., Oh, W., Ehrenkranz, R. A., Shankaran, S., Donovan, E. F., Wright, L. L., Lemons, J. A., Tyson, J. E., Korones, S. B., Bauer, C. R., and Stoll, B. J. (2000). Sex differences in outcomes of very low birthweight infants: the newborn male disadvantage. *Arch Dis Child Fetal Neonatal*, 83:F182–F185.

Sá, M. I., Fonte, M., Carvalho, C., Soares, P., Almeida, A., Januário, A., Gouveia, S., and Saraiva, J. (2012a). Premature infants under 27 weeks gestacional age: predicting outcome and improving parental counseling. *to be submitted*.

Sá, M. I., Fonte, M., Saraiva, J., Carvalho, C., Soares, P., Januário, A., Gouveia, S., and Almeida, A. (2012b). Management and outcome of extremely premature infants. In *Book of Abstracts of XLI Jornadas Nacionais de Neonatologia da Sociedade Portuguesa de Pediatria*, pages 46–47. Secção de Neonatologia da Sociedade Portuguesa de Pediatria.

Sá, M. I., Fonte, M., Saraiva, J., Carvalho, C., Soares, P., Januário, A., Gouveia, S., and Almeida, A. (2012c). Management and outcome of extremely premature infants. In *Book of Abstracts of 17th World Congress on Controversies in Obstetrics, Gynecology & Infertility (COGI)*, page xx. Federação das Sociedades Portuguesas de Obstetrícia e Ginecologia (FSPOG).

The Express Group Members (2010). One-year survival of extremey preterm infants after active perinatal care in sweden. *Journal of the American Medical Association*, 301(21):2225–2233.

Turkman, M. A. A. and Silva, G. L. (2000). *Modelos Lineares Generalizados da Teoria à prática*. Edições SPE, Lisboa.

Tyson, J. E., Parikh, N. A., Langer, J., Green, C., and Higgins, R. D. (2008). Intensive care for extremely prematurity-moving beyong gestational age. *N Engl J Med,*, 358(16):1672–1681.

Vittinghoff, E. and McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemology*, 165(6):710–718.

Wang, Y., Tanbo, T., Ellingsen, L., Abyholm, T., and Henriksen, T. (2011). Effect of pregestational maternal, obstetric and perinatal factors on neonatal outcome in extreme prematurity. *Arch Gynecol Obstet.*, 284(6):1381–1387.

Waters, T. P. and Mercer, B. M. (2009). The management of preterm premature rupture of the membranes near the limit of fetal viability. *American Journal of Obstetrics & Gynecology*, pages 230–240.

Whitford, A. B. (2005). *Encyclopedia of Social Measurement*, volume 1, chapter Correlations. Elsevier.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining:Pratical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3 edition.

Worth, A. P. and Cronin, M. T. (2003). The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for humam health effects. *Journal of Molecular Structure (Teochem)*, pages 97–111.

Zeitlin, J., S.Draper, E., Kollée, L., Milligan, D., Boerch, K., Agostino, R., Gorther, L., Reempts, P. V., Chabernaud, J.-L., Gadzinowski, J., Bréart, G., Papiernik, E., and the MOSAIC research group (2008). Differences in rates and short-term outcome of live births before 32 weeks of gestation in europe in 2003: Results from the mosaic cohort. *Pediatrics*, 121(4):e936–e944.

Zwanenburg, A., Meijer, E., Jennekens, W., van Pul, C., Kramer, B., and Andriessen, P. (2012). Automatic detection of burst synchrony in preterm infants. In *Book of Abstracts of 34th Annual International Conference of the IEEE EMBS*, pages 4720–4723.